Modelling land use using demographic forecasting and local optimisation:

A case study of general education provision in

Riyadh, Saudi Arabia.

Salman Altuwariki

A dissertation submitted in partial fulfilment

Of the requirements for the degree of

Doctor of Philosophy

Of

University College London

The Bartlett Centre for Advanced Spatial Analysis

University College London

25 March 2023

Declaration

I, Salman Altuwariki confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Globally accepted guidelines for land use allocation in Riyadh, Saudi Arabia have been based on an outmoded practice that was created over a century ago. This approach is based on a mix of predetermined population densities, walking distances, and per person area ratios. The latter criterion is essentially based on a worldwide average for facility areas and user numbers. The fundamental criticism levelled at such practices is their insensitivity to population trends and limited land resources. In this context, this research is aimed at updating common practice in the light of population growth and residential mobility projections at the city and district levels. The models introduced aim to provide comprehensive and adaptable simulation tools for optimising any type of land use provision standard over a specified time period. The simulation environment makes use of an agent-based framework that adapts and integrates a number of well-known methodologies, including Cohort Component Modelling (CCM) for population projection, Spatial Interaction (SI) modelling for residential mobility, and AutoRegressive Integrated Moving Average (ARIMA) for various ratio extrapolation. Additionally, new hybrid concepts and approaches have been evaluated, including a household based CCM and the use of Neural Network algorithms (NN) to forecast residential mobility. The case study focuses on Saudi populations in Riyadh, Saudi Arabia where the three general education stages at elementary, middle, and secondary levels were optimised for both genders. Moreover, the optimisation time horizon spans 50 years, from 2020 to 2070 while the focus of research at the city level optimises the conventional ratio of area per student based on the present stock of education allocated land and a land consumption ratio defined for every five years. The district level optimisation, on the other hand, balances the demand and supply of education over 50 years by utilising the Ministry of Education's (MOE) predesigned school prototypes. The research findings demonstrate the feasibility of developing a tool for optimising land use guidelines that is capable of producing acceptable outcomes while being sensitive to demographic change and land resource availability.

Impact Statement

Because the field of optimising land use provision is lacking, it is anticipated that this study will have an influence inside and outside the academic field. This research offers a novel strategy for optimising the provision criteria for land use in urban planning by concentrating on population demography, small areas, pre-designed facilities, and specific allotted lands. In addition, the tool that was developed as a result of this research goes beyond the merely theoretical development of techniques and is, rather, in a state where it is ready to be used, as demonstrated by the presented case study. Furthermore, the method has been extensively documented for the purposes of education, replication, development, and wider adoption. Moreover, the model is essentially generic and modular in order to extend its applicability beyond the optimisation of provision standards and serves as a complete urban simulation system. This was done to ensure that the model could be used as a test platform for carrying out tasks related to both explanatory and forecasting aspects. Concerning the model's application and its results in Riyadh city, this study expands current practise of standards calibration by factoring in household-based population growth predictions, residential mobility forecasts, and the usage of pre-designed facilities to optimise the availability of services. These three elements lead to a better degree of consistency across the various authorities, whose existing practises are contradictory to some extent. In addition, to the best of the author’s knowledge and experience, the Saudi Arabian planning authorities have not yet put into action a comprehensive urban simulation model that is operative at the district level. In addition, this study gives statistical data insights that encompass its availability, inconsistency, and usability, which finally will lead to a stronger understanding of the data limitations and possible curation methods. On the business side of things, this research might help improve the district level population profiling studies, which current commercial businesses could use to evaluate the demand for their activities. Also, it could be helpful in the evaluation of privatising services that are currently provided by governmental bodies.

Table of Contents

List of Tables

List of Figures

Acronyms

Age Specific Complete Fertility rates (AS-CFR)

Age Specific Fertility Rates (AS-FR)

Age Specific Total Fertility rates (AS-TFR)

Agent-Based Modelling (ABM)

ArRiyadh Development Authority (ADA)

Artificial Neural Networks (ANN)

AutoRegressive Integrated Moving Average (ARIMA)

Cohort Component Modelling (CCM)

Completed Fertility Rate (CFR)

Conditional Probability (CP)

Data Quality (DQ)

Deep Learning algorithms (NN)

Deterministic Reweighting (RD)

Discrete Event (DE)

Floor Area Ratio (FAR)

General Authority for Statistics (GAS)

Genetic Algorithms (GA)

Hyperbolic Tangent (Tanh)

Integrated Development Environment (IDE)

Iterative Proportional Fitting (IPF)

King Abdulaziz City for Science and Technology (KACST)

Machine Learning (ML)

Maximum Building Footprint Area ratio (MBFAR)

Mean Absolute Deviation (MAD)

Mean Average Percentage Error (MAPE ± SD)

Metropolitan Development Strategy for Riyadh (MEDSTAR)

Microsimulation (MS)

Ministry of Education's (MOE)

Ministry of Health (MOH)

Ministry of Justice (MOJ)

Overview, Design concepts, and Details (ODD)
Planning Support System (PSS)
Process Modelling Library (PML)
Rectified Linear Unit (ReLU)
Riyadh Municipality (RMU)
Scatter Search (SS)
Simulated Annealing (SA)
Spatial Interaction (SI)
Standard Deviation (SD)
System Dynamics (SD)
Tabu Search (TS)
The Ministry of Municipal and Rural Affairs (MOMRA)
The United Nations Department of Economic and Social Affairs (UN-DESA)
Total Fertility Rates (TFR)
United Nation Environment Programme (UNEP)

Chapter 1
Introduction

The Kingdom of Saudi Arabia has rapidly developed in the period between the 1960s and the 1980s (El Mallakh and El Mallakh, 1982) while its capital Riyadh city is considered to be one of the fastest growing cities in the world (ADA, 2011). In addition, the Kingdom's population has been rapidly growing and it is projected to increase from around 27 million in 2010 to 40 million in 2050 even though the growth rate will decline from around 2% to 0.5% per annum for the same period (UN-DESA, 2013). Rapid urban development and population growth were the main catalysts for the development of the urban planning system and its policies since the Kingdom was established and this is reflected in the current land-use planning paradigm. At the Kingdom level, the current land-use planning system is mainly based on ensuring maximal land use coverage, appropriate population densities and area per person in determining the public facilities and services needed that is guided by the Ministry of Municipal and Rural Affairs (MOMRA)[1]. However, some cities have their own planning authorities such as ArRiyadh Development Authority (ADA)[2,3] which can alter the use criteria for land use provision. Despite the availability of these guides some related governmental bodies such as the Ministry of Education (MOE) follow different approaches to provide facilities such as having pre-designed school types that is constraining its facilities provision. Nevertheless, given the shortage in supply in most public facilities and services highlighted by the ADA action plan for coordination and provision of public services in Riyadh (ADA, 2011). The ADA has developed a location-allocation model to tackle this issue by considering the same criteria in MOMRA guidelines in addition to the population growth over the long-term. The ADA updated standards are designed to fulfil the shortage by lowering the standards mainly by providing one standard per service for all districts despite the variations among the districts in terms of land resources and population demography and

[1] Recently renamed to Ministry of Municipal and Rural Affairs and Housing (MOMRAH)
[2] The name of Riyadh city can also be written as ArRiyadh.
[3] The Royal Commission for Riyadh City (CRCR) is the new name for ArRiyadh Development Authority, which was formerly known as The High Commission for the Development of ArRiyadh (HCDA). This planning body will primarily be referred to as ADA. However, the references abbreviations will appear after the name that is printed on the referred document.

the standards do not utilise different building pre-designs available for certain services. This problem extends to both the public and private sector as they are based on the same planning approaches in providing services. Thus, the ability to test an optimised fine grain land-use planning model-based system could improve the government performance, private sector profitability and the individual life qualities. However, the methods and approaches for optimising land use provision are complex, not yet well developed and require the use of new forms of Planning Support System (PSS) which we will develop here in this thesis.

The general education system in Saudi Arabia is formed of 3 stages which are Elementary, Middle and Secondary schools. The Elementary school accepts children when they are 6 years old and its interval extends for 6 years, while both the Middle and Secondary schools are each 3 years long. Therefore, the range of age groups in this education system covers is between 6 and 18 years old and spreads over 12 years. Moreover, the educational system adopts mainly a total separation between genders. Therefore, typically separate buildings are provided. The main type in education is the morning schools for the earlier mentioned age cohorts. However, other types for these stages exist such as night schools, disability-oriented schools, literacy schools, international schools, and home study. These types of schooling accept different age cohorts and have special requirements. For example, the home study starts with the Middle school stage and the students need only to attend their exams at the end of each semester. Another example is that international schools are normally mixed gender schools at the Elementary stage. Moreover, schools in Saudi Arabia can be either under the authority of the government or the private sector. The schools' acceptance boundaries for students in Riyadh city are softly drawn for the schools' directors by the MOE. In other words, parents are not limited to certain schools based on their place of residence as the place of work might be their criterion for choosing a school. However, the schools' directors have a list of the district priorities for their applicants to ensure enough places available for nearby residences which are within the school capacity.

1.1 The Research Problem

The total number of general education schools increased from 2039 in 2010 (HCDA, 2010) to around 3346 in 2017 (GAS, 2017) of which 46% are provided by the private sector as the main contributor to this rapid growth. However, the growth in the number of schools is dependent on rented buildings and these also comprise around 61% in the private sector (GAS, 2017). The schools under the government authority are in a better position as the percentage of rented buildings is about 9% for girls' schools and 14% for boys' schools in Riyadh city (GAS, 2017). The estimated land area required to satisfy its allocation standards consists of 61% of the total land dedicated to public facilities in Riyadh city (HCDA, 2010). This high investment from both sectors must be guided to provide optimal supply with high sensitivity to the land resources consumed.

Planning with the use of rules of thumb and averages in land allocation cannot be viewed as an optimal solution through the application of standards such as Smart Codes and local participation which form the basis of a PSS. The value of the land and the expected disposable income of various economic population profiles can contribute to suggesting optimal land allocation according to the well-known supply and demand mechanisms that would work to prevent any unbalancing in these factors. Moreover, if the forecasting of population growth and its economic feasibility becomes more integrated into planning, then investments related to non-linear demand can be met with the best scenarios for current and long-term sufficiency. A PSS that utilises Agent-Based Modelling (ABM) is a strong means of linking the not yet fully developed land allocation process with population growth and economic feasibility. There must be a clear framework at least for these two elements to be incorporated in virtual simulations which generate an optimal disaggregated land use allotment related to the population demography.

Planning in Riyadh continues to adhere to the fixed land use standards of the early planners, but with substantial public and stakeholders' involvement. Moreover, even the latest trials in enhancing the available codes followed comparison with global averages that are not suitable for the city demographics. Furthermore, minimising the standards seems to be the general objective. Given the estimated amount of land and the acquisition cost necessary to provide what is perceived as optimal provision, the observed decline in growth rates and the current economic challenges with low oil prices do not make this feasible. Additionally, the current procedures could be improved to align with the Saudi Vision 2030 for effective operational governance and improved living standards. This research thus provides an alternative solution to the urgent problem facing Saudi Arabia in the whole process of land allotment.

1.2 Identifying the Gap

In the past, land area allocation standards at the urban planning level were proposed based on global averages, neighbourhood density, and walking distance. One problem is that this approach is not sensitive to national or local requirements. For example, in countries with an ageing population, the area allocated for general education would be oversupplied as time progresses but any contemporary approach should balance demand and supply. The techniques for optimising land use are complex, data-intensive, and call for subject matter specialists. The need for a comprehensive tool with high transparency and adaptability to boost planner engagement and the planning process as a whole is therefore significant. This is what we aim to provide here.

1.3 Contribution

The allocation area in the models we propose are based on the possible population growth and residential mobility scenarios rather than global averages. This method is expected to provide different insights which could lead to increased accuracy and efficiency in land allocation over longer time horizons for the urban planning process. In this study, we will propose a PSS that is comprehensive in considering urban factors, fundamentally generic to be used in different cities, optimise land usage at the city and district level, and deal with conventional standards (i.e. area per person) and facility-specific standards that use pre-designed building prototypes.

1.4 Research Aims and Objectives

To develop an urban simulation model for Riyadh city, we need to be able to project population change at the district level using different scenarios based on zonal growth factors. Also, we need to estimate the demand for educational facilities over different time horizons and then optimise the schools' provision standards given the constraints of available 'white lands', that is land which is available for new uses.

1.5 Substantive Importance of the Study

The importance of this research lies in identifying the gap in the current practice of function allocations by urban planners from a demographic and economic perspective. It aims to develop a generic methodology that integrates these two aspects in function allocation at the proposal level of the urban development process over the short and long term. Moreover, it identifies the most influential data required for this kind of study in planning and thus aims to promote such data collection by the different authorities. This study offers a general optimisation paradigm, of which the provision of education facilities is one illustration.

Chapter 2

Literature Review

There are three main sub-sections in this chapter. The city of Riyadh is discussed in the first sub-section in terms of its population demographics, stages of master planning, recognition of general sub-optimality in the provision of public facilities, and standards that are implemented for educational provision. The second subsection will go over the broader issue in land use planning that results in inefficient allocation. This section outlines important urban planning strategies using historical examples. The third sub-section focuses on alternative methods for resolving the land use allocation problem, which come from disciplines in the social sciences and economics and are loosely related to PSS. This section also provides a brief historical overview of the evolution of these systems. Furthermore, it provides two examples of both early and modern comprehensive models while describing the fundamentals of creating comprehensive urban models. The final step will be the demonstration of three models that are related to the wider context of issues providing public amenities and commercial activities in Riyadh city.

2.1 The case study: Riyadh, SA

Over the course of the past fifty years, Riyadh has expanded from a small town with fewer than a half million people to a sprawling city with more than six million residents. Moreover, there are hardly many examples anywhere in the world that can compare to the size and pace of its evolution (Al-Hathloul, 2017). The population demographics and development of the city, as well as its land use provision standards, are discussed in this section.

2.1.1 Riyadh's population

Riyadh's population numbers jumped from approximately 1.3 million in 1987 to 6.5 million in 2016, despite the decline in the total annual growth rate from nearly 9% to 4% for the same period. The population age-group under 15 years consisted of approximately 26.5% of the population in 2016 down from 38.5% in 1987, while family size fell somewhat from 6.23 to 5.7 over the same period (HCDA, 2016b). The calibration of provision standards must take into account such demographic shifts, particularly over the long term. Furthermore, the age structure of the Saudi population is significantly different from non-Saudis, see Figure 1 below for a 2016 comparison.

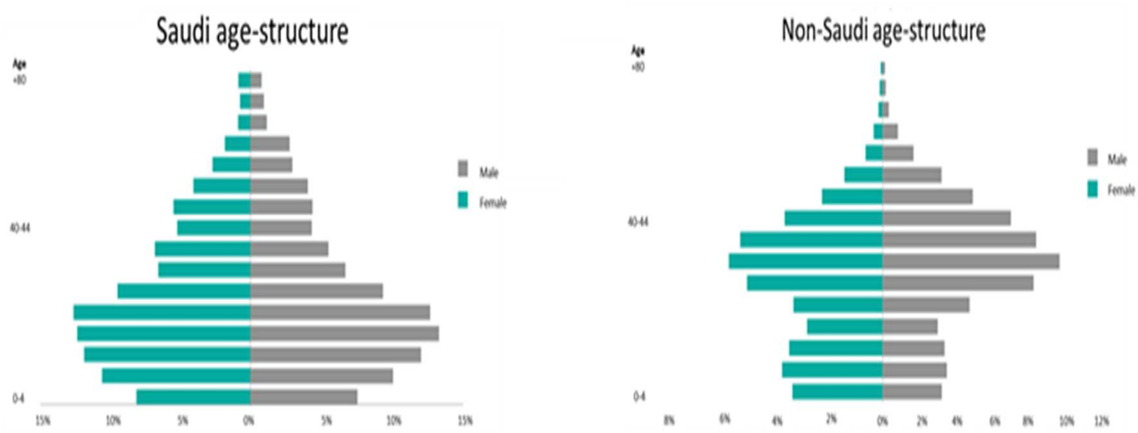


Figure 1. The population structure of Saudis and Non-Saudis in Riyadh 2016 Adapted from (HCDA, 2016b)

In 2016, immigrants consisted of approximately 30% of the Saudi population (GSA, 2016). However, immigrants in Saudi Arabia have no right to stay after the period defined in their contracts. Therefore, the demographic changes are related to the (in and out) immigration rates more than their fertility and mortality rates, so immigrants must be considered a special population category when planning for their needs. In this research, the immigrant population was not included for optimising the schools provision standards given their special case.

2.1.2 Riyadh's planning history

The 1930s walled city of Riyadh was less than 1 km^{2} (Al-Hathloul, 2017), but with the city's rapid growth, by 1977 the built up area reached around 73 km^{2}, and has increased seven times from 1983 to 511 km^{2} (Daghistani, 1985). Currently, the spatial coverage defining its urban limits starts from 2395 km^{2} and reaches 5961 km^{2} for the urban environs limits (Qhtani and Al Fassam, 2011). Figure 2 presents Riyadh's different phases of urban growth between 1900 and 2000. In the early stages, Riyadh's growth was not a planned development, but triggered by the construction of royal complexes (referred to as Qasar in the figure) and scattered districts (Al-Olet, 2004).

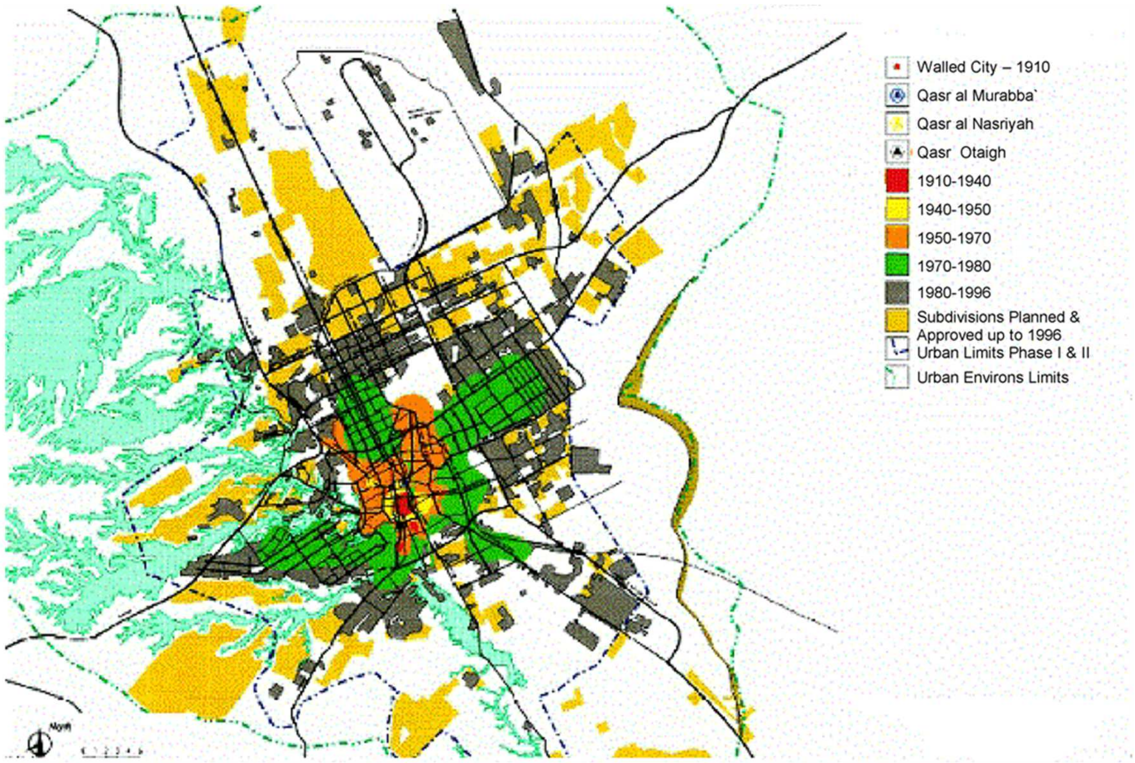


Figure 2 Riyadh phases of urban growth (1900 – 2000), adapted from (Al-Hathloul, 2017).

Currently, most land development in Saudi cities follows formal land subdivisions (Abdulaal, 2012). Moreover, a unified subdivision regulation guide has been imposed by the MOMRA for conventional land subdivisions planning (Qurnfulah, 2015). However, some cities have a special planning body, like the ADA in Riyadh city, which controls the planning process in great detail. Riyadh city has been through three phases of master planning from the first consultants Doxiadis, to SCET, and then the ADA respectively (Al-Hathloul, 2017). Despite the inherent sub-optimality of land-use allocation practiced globally, this problem was recognized by the official bodies in each phase of Riyadh city master planning.

2.1.2.1 The Doxiadis masterplan

The first town planning office in Saudi Arabia was under the Municipal Affairs in the Ministry of Interior, which commissioned the planning firm Doxiadis to develop Riyadh city's first master plan between 1968 and 1972 (Daghistani, 1985; Middleton, 2009). As we show in Figure 3, The first approved master plan For Riyadh city, adapted from (Middleton, 2009), was the Doxiadis plan which was based on a commercial spine orthogonal to a sub-commercial axis serving a grid-based plan of 2 km x 2 km neighbourhood units.

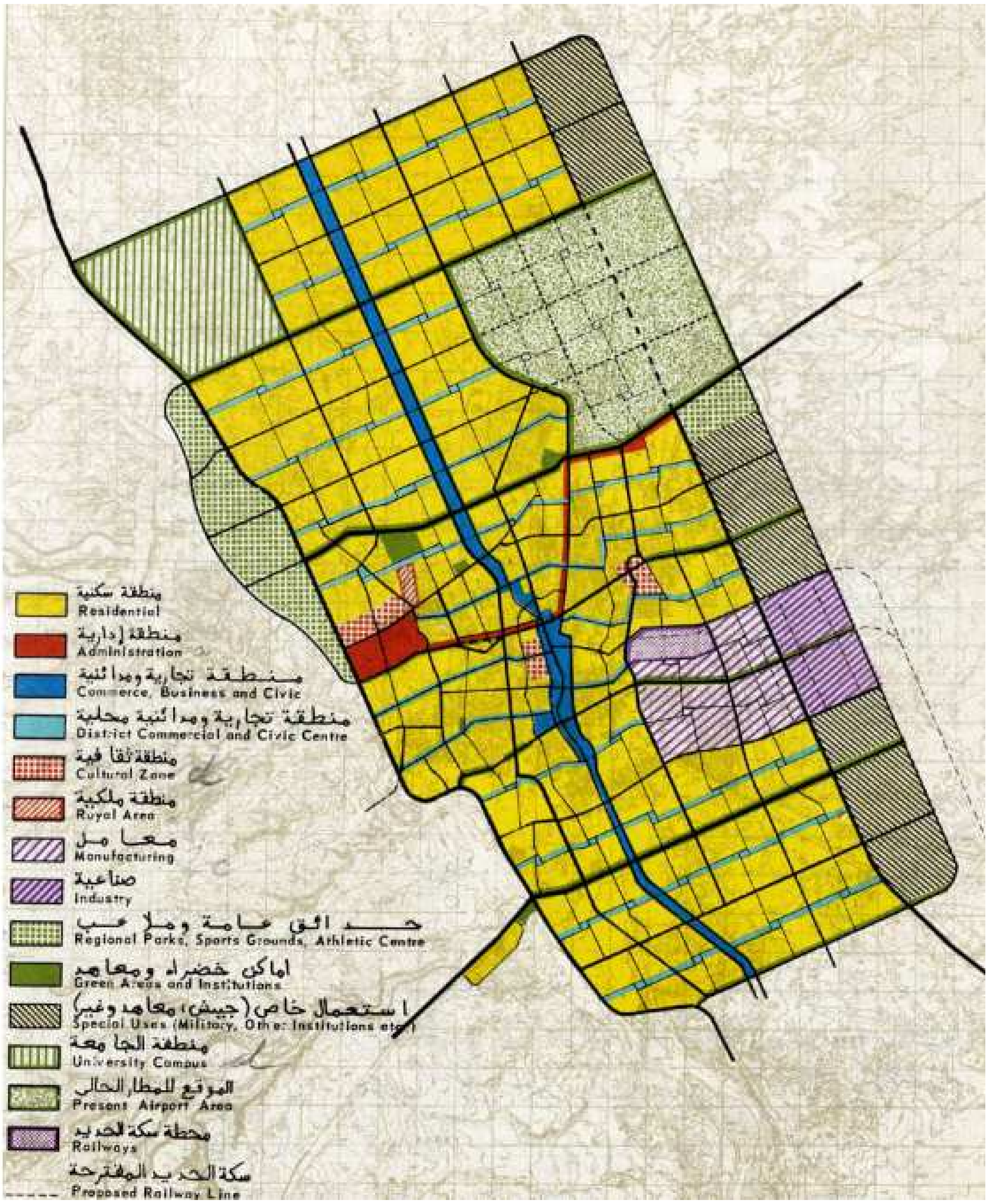


Figure 3 The first approved master plan For Riyadh city adapted from Middleton (2009).

According to Daghistani (1985), the first master plan's objectives were: 1) to calculate the potential expansion of the city by analysing its economic, administrative, educational, and other activities; 2) to estimate the size of the city in the years 1975, 1980, and 2000 in terms of population and space requirements; 3) to specify the location of its primary operations, including civic and commercial areas, administration, industry, education, health, sports, recreation and etc; 4) to provide a final plan that will guide and regulate the city's future expansion; 5) to anticipate and divide the development phases for the given target years; and 6) to provide a summary of the development programme up to the year 2000 and the immediate development programmes for the first two five-year periods up to 1980. Clearly the plan's aims were focused on the overall potential growth of the city for major functions such residential, commercial, and educational uses. Moreover, it was supplemented with further details about expected population size in each planning unit and their requirements for public and commercial services. See Table 1 for the Doxiadis hierarchy of functions based on the community classes in the masterplan adapted from Middleton (2009)

Table 1 The Doxiadis hierarchy of functions based on the community classes in the masterplan adapted from Middleton (2009).

Hierarchy of Communities for Riyadh Kingdom of Saudi Arabia, 1972

Order of Community	**Population Size**	**Functions**									
		Education	**Commerce**	**Business**	**Services**	**Health**	**Transport**	**Socio-Cultural**	**Recreation**	**Industrial**	**Administrative**
Class I	40								Small Square		
Class II	250	Kindergarten/ Primary school	Corner Kiosk						Play Area for Pre-school age		
Class III	1,500	Kindergarten/ Primary /secondary school	Small Local market		Barber		Bus-stop Post Office	Small Mosque	Play Area for pre-school age children		
Class IV (Neighborhood or Small city)	9,000	Secondary school	Small shopping center	Real Estate Office	Gasoline Station Plumber, electrician. Photographer	Dispensary Doctor's Office	Taxi station Small Post-Office Telegraph Office	Mosque, Local Club	Small Public garden square	Handicrafts	Police Station for every 4-5 communities
Class V (City)	50,000	Secondary/Trade Commercial school	Central Market	Bank Insurance Offices	Lawyers Doctors Engineers	Health Center, Private Clinics	Local inter-urban Bus Railway Airport Central Post-office	Large Mosque, Meeting Hall, Library	Playing Fields, Small Parks	Light Industries, Intermediate Industries	Local Police, Division Municipality Offices
Class VI (Large City)	300,000	Small University	Commercial Center	Large Bank, Headquarters business firms Large representative offices	Professional Organizations Large local offices	General Hospital	Large Local and inter-urban Bus Station	Central Mosque, Museum, Art galleries, Small Concert Hall	Stadia, Parks, Theatres	Intermediate and Heavy Industry	Government Headquarters
Class VII (Metropolis)	2,000,000	Large University	Large Commercial Center Wholesale center/ warehousing	Central Offices Banks, Insurance, Headquarters large firms	Headquarters of large private offices	Large General Hospital Specialized Hospital	Central Rail, Bus Stations, Large Airport	National Museums, Permanent Art Galleries, Exhibition Halls	Olympic Stadium, Race course Large Parks	Heavy Industry, Special Industries.	Seat of National Government

(Dox-SAU- A19 pg. 263 Table 47)

Table 1 shows that the planned communities in the Doxiadis masterplan were classified into 7 classes, each of which has an expected population size and services to be provided. These criteria for the provision of services are common practice in the early days of urban planning as will be discussed on page 35. However, early on in the implementation of the Doxiadis plan, it was deemed unsuccessful due to the leapfrog development that happened. The term leapfrog development refers to the development of land that requires public facilities in advance of the masterplan's development timeline (Al-Olet, 2004; Daghistani, 1985). This type of expansion was brought on by a number of factors, including: 1) the mid-1970s boom in the economy and government expenditure, which coincided with a lack of capacity to control urban growth; 2) Doxiadis' underestimation of the predicted 30-year growth and; 3) the absence of detailed guidance and instructions for implementing the first master plan (Daghistani, 1985). In addition, according to Daghistani (1985), the plan principle of hierarchical communities was not fully implemented due to the underestimation of the attraction of commercial activities to roads and the absence of an intermediate level of services between the city (Class V) and neighbourhood levels (Class IV). As a result, the community centres were not produced as originally intended. Daghistani clearly stated that:

> *'Indeed the whole question of the distribution of services throughout the city — such an important part of the Master Plan approach — simply did not happen' (Daghistani, 1985, p. 162).*

In response to the deficiencies of the initial masterplan, a second attempt was made.

2.1.2.2 The SCET masterplan

In 1975, the physical planning responsibilities were transferred to the newly established MOMRA (Daghistani, 1985), which commissioned the French planning firm SECT International in 1976 to review and revise the Doxiadis plan which had been overtaken by events (Daghistani, 1985; Mubarak, 2004). The approach in the second plan was dramatically different, with its focus on implementation and action plans at different levels, as follows: 1) the master plan defined the city overall policies and structure; 2) it produced a city level 5-year action plans co-ordinated between the different government agencies; and 3) it prepared detailed small area action plans (Daghistani, 1985). Therefore, the SECT plan was not designed to place limits on Riyadh's expansion; rather, it served as an organisational tool and provided rules for how Riyadh's physical expansion should occur (Al-Hathloul, 2017). Moreover, providing the adequate level of public services to the resident population was one of the main objectives of the small area action plans (Daghistani, 1985). However, the provision of the adequate level of services at the small area action plan level with the coordination between government agencies did not sustain. This issue was evident in the reports produced as part of the third master planning effort conducted by ADA.

2.1.2.3 ADA masterplan

The third attempt to control the city's planning came in 2003 with the introduction of a Metropolitan Development Strategy for Riyadh (MEDSTAR) (Al Naim, 2013). This comprehensive plan was introduced by the ADA, which is the latest executive, technical, and administrative planning body for Riyadh, established in 1983 (HCDA, 2015). MEDSTAR aims to control the city's rapid urban growth and address the city's physical, environmental, and potential development problems (Al Naim, 2013). Figure 4 shows the MEDSTAR 2030 master plan for Riyadh city.

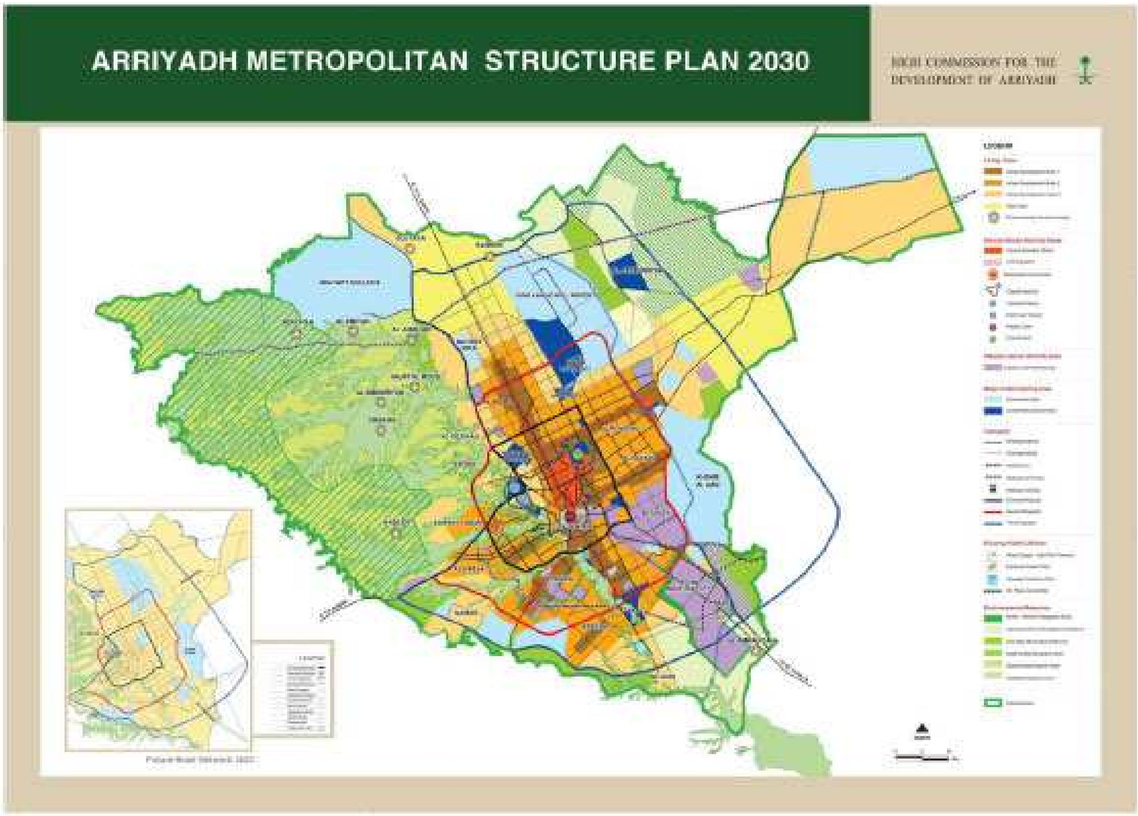


Figure 4. MEDSTAR 2030 master plan for Riyadh city adapted from (Qhtani and Al Fassam, 2011) .

According to Middleton (2009), at the city level, the plan focused on the construction of several high density commercial sub-centres, the redevelopment of the historical city centre, and controlling growth within physical limits and timed phases. At the residential community level, there was a reconsideration of the Doxiadis 2 km x 2 km grids to be made four times smaller in the future with a population of approximately 3000 to 4000. Also, the plan experimented with urban forms other than the usual unit of the Doxiadis neighbourhood attempting to enhance responsiveness to social and environmental demands. The comprehensive plan addresses a wide range of subject matter facets and produces research reports, technological solutions, and executive action plans (Qhtani and Al Fassam, 2011), some of which are centred on the distribution of public facilities and commercial services. Moreover, MEDSTAR had periodic updates and was most recently revised in 2011 (ADA, 2015).

Currently, the supply of urban functions in Saudi Arabia is based on a hierarchical community size structure led by MOMRA. This approach is essentially associated with maximal coverage, population thresholds, and land area per person for assigning public and commercial lands (MOMRA, 2015). However, these basic guidelines might be adjusted by the particular planning arms in the different cities in Saudi Arabia according to their assessment of local needs. In Riyadh city, ADA has been the agency championing these rules. Despite the presence of this framework, it can be seen as nominal. In conventional land use planning; for instance, 33% of the total developable area is dedicated to public services and roads. However, the specific distribution for the various functions is dependent upon the design speculations of the planners and the permission of the municipal officials (Alskait, 2003). In the parts that follow, the ADA action plan for the state of education provision and its response to its outcomes, as well as the MOMRA's contribution to directing the supply of standards, will be explained. Additionally, the provision of commercial activities will be touched upon to highlight the similarities in the various approaches to setting the standards.

2.1.3 Current standards and the shortage level in education

The ADA (2011) report about the public services status in schools as one of the MEDSTAR studies, indicates a significant shortage in all public functions expect for the provision of mosques. Moreover, the 110 km^2 stock of public land is insufficient, with a need to allocate an additional 32 km^2 of land in Riyadh to satisfy its need for public facilities up to the year 2029. Furthermore, the most affected functions are educational, and they consist of approximately 61% of the total land needed to satisfy the supply of public functions. These results are calculated after adjusting the land allotment standards to a lower level of provision through a comparison with the standards of other national, regional, and global cities (ADA, 2011). Moreover, the latest revision for the standards known to the author was in 2016. Table 2 provides the summary of the education facilities shortage and a comparison between the 2009 and 2016 ADA standards with the latest standards suggested by MOMRA.

Table 2. Summary of the public facilities shortage and a comparison between the old and new ADA standards with the latest MOMRAH standards

Education stage	Criteria	MOMRA (2015)[1] standards	ADA (2009)[2] standards	ADA (2016)[2] standards	Shortage in (2016)[2] using latest ADA standards
Elementary	Catchment radius (m)	500 B	550 B	550 B	-487 M -537 F
	Population threshold	3200-4000 B	6000 B	8000 M 8500 F	
	m^2/person	15-20 B	1.3 M 1 F	1 M 0.94 F	
Middle	Catchment radius (m)	1000 B	800 B	700 M 800 F	-245 M -241 F
	Population threshold	6000-8000 B	12000 B	16000 M 18000 F	
	m^2/person	20-25 B	0.7 M 0.6 F	0.65 M 0.57 F	
Secondary	Catchment radius (m)	2000 B	1500 B	800 M 900 F	-212 M -163 F
	Population threshold	11000-13000 B	18000 B	20000 M 25000 F	
	m^2/person	25-30 B	0.9 B	0.6 M 0.5 F	

B Both male and female
M Male students
F Female students
1: (MOMRA, 2015)
2: (HCDA, 2016a)

Table 2 shows the standards' three conventional criteria of Catchment radius, Population threshold and the area per person for representing the standards. When comparing the old and new ADA standards, it can be noted that by large, the standards have been reduced and some distinctions between males and females have been introduced. Overall, the shortage in year 2016 required the construction of more than 1500 schools. However, when we compare the standards of MOMRA to ADA, a significant variation between the area per person standard can be noted. This is suspectedly due to the difference in the method followed to calculate the standards as the net buildings footprint area or gross land area. Another method for the standardization could use pre-designed building types for the different education stages that are based on optimal operational capacities. The latter type of standard is adopted by the MOE and adds to the complexity of comparing the standards. Table 35 illustrates the different types of pre-designed school types by education stage, footprint area and students capacity.

According to ADA (2011), there are several major problems facing the public facilities: there are 1) the transformation of the facilities' dedicated lands to other uses; 2) the high percentage of rented buildings for the facilities; 3) the shortage in land owned by authorities; 4) the significant difficulty in distributing and determining the size for facilities; 5) the need for unified supply criteria for the facilities; 6) the weak communication between the different authorities; 7) the lack of spatial data about the facilities; and 8) understanding the effects of implementing new services, like E-government.

On the other hand, the standards for the provision of commercial activities are less detailed than the standards of public facilities in terms of types and criteria. First, no information is available about the ADA 'Commercial activities distribution study', which is part of the MEDSTAR masterplan that could be referenced in this matter. Second the 2015 MOMRA standards abandoned the major classification of the types of commercial functions: consumer goods, durable goods, professional services, and light industrial and warehouses activities. See Table 3 for a summary of the 2015 MOMRA standards of commercial activities. Moreover, It should be noted that the approach for the development of commercial functions is based on a comparison with the standards of other global cities (MOMRA, 2015). Due to weak control over the public facilities, which includes more detailed standards, we can assume that the commercial distribution of function is sub-optimal.

Table 3 Summary of the 2015 MOMRA standards of commercial activities (MOMRA, 2015).

Criteria	**City Level**	**Planning Hierarchy**		
		Neighbourhood	**District**	**City**
Population Size (Thousands)	**small city**	3-4	10-15	less than 100
	medium city	4-5	10-20	100-200
	large city	5-6	15-20	larger than 200
Number of Shops (Per Person in Thousands)	**small city**	5-10	5-10	5-10
	medium city	10-15	5-10	5-10
	large city	10-15	10-15	10-15
Area Per Person (m^2)	**small city**	0.25-0.5	0.25-0.5	0.5-1
	medium city	0.5-1	0.5-0.75	1-2
	large city	1-1.5	0.75-1	1.5-3
Land allocation (m^2)	**small city**	750-2000	2500-5000	-
	medium city	2000-5000	5000-10000	-
	large city	5000-9000	10000-15000	-

Table 3 shows similar allocation criteria to the provision of education facilities as well as the additional classification of city level and the planning hierarchy within the cities.

However, these standards remain at the same level of the early planning attempts to set provision guidelines for these commercial activities.

2.1.4 The weaknesses of current planning practice

The use of global averages in developing standards is not sensitive to population demographic requirements. For example, the 2005 planning guide for public facilities by MOMRA defines the supply of elementary schools at 6.25% of the total population at maximum. In the updated 2015 guide, the percentage is increased to 8.5% at the maximum instead of 6.25%. However, the current percentage of that age group exceeds 10 %, which does not only cause a spatial allocation problem, but also extends to the learning environment due to shortages in supply produced by current standards. Another example could be drawn from the population composition in some districts, which are mainly composed of international immigrants with none or a small number of children. Therefore, their need for education facilities would be less than the districts housing indigenous citizens. Alskait (2003) highlights another problem that could be caused by the land subdivision regulations that allows owners to subdivide their plots and increase population density by up to 50%. This increase in densities is not accounted for in these regulations. Therefore, lowering the standards could mean lower quality of living. As we will see in the following section, these issues are fundamentally caused by the way land use planning was first established and has remained unchanged since its start.

2.2 Inclusivity and Optimality in Land-use Planning

Land-use planning is not only about the general location, distribution, and extent of different uses, but also inherently includes aspects, such as social, economic, and environmental objectives which seek optimality or sustainability. Land-use planning is essentially the balancing tool for competing participants and approaches (Pickardt and Wehrmann, 2011). Regardless of the origin of land-use planning be it as a tool for agricultural, rural, or city planning, land-use planning has adopted these objectives. For example, in the United States, both agricultural policies and city planning considered economic and social adjustments to the urban and rural system as early as the 1930s and 1940s (Akimoto, 2009).

Another example is from United Nation Environment Programme (UNEP) whose definition of land-use planning states:

> *'Land use planning is a systematic and iterative procedure carried out in order to create an enabling environment for sustainable development of land resources which meets people's needs and demands. It assesses the physical, socio-economic, institutional and legal potentials and constraints with respect to an optimal and sustainable use of land resources, and empowers people to make decisions about how to allocate those resources' (UNEP, 1999, p. 14).*

The UNEP definition highlights the ideas of people's needs, assessment of optimality for various aspects, and legal limitations. Moreover, the Canadian Institute of Planners (CIoP) defined planning as follows:

> *'planning means the scientific, aesthetic, and orderly disposition of land, resources, facilities and services with a view to securing the physical, economic and social efficiency, health and well-being of urban and rural communities' (CIoP, 2013).*

The Canadian view of land-use planning comprehensively includes additional aspects, such as aesthetics, position of facilities and services, health, and well-being. These definitions demonstrate clearly that land-use planning is a process of optimising various aspects for the benefit of all stakeholders. The land use allocation standards used by urban planners have been used as a tool to generate optimality in urban planning. However, despite its importance these land use provision standards remain overlooked right after their first conceptualisations until the present as we will demonstrate in the next subsection.

2.2.1 The Allocation Standards

This section explains the urban planners' approaches to formalise the land use provision standard, these approaches are divided in a temporal manner into early and current approaches, which then will be followed by concluding remarks.

2.2.1.1 The urban planners approach.

A. Early urban planning

In urban planning, one of the earliest perceptions on how land allocation might be planned optimally was proposed by Harland Bartholomew (Akimoto, 2009). Bartholomew analysed many American cities and suggested the relationship between population size and the amount of area needed for various uses at an aggregated level such as areas for residential, public, commercial, industrial etc. (Bartholomew, 1932). For example, schools were considered as part of the public and semi-public areas in self-contained cities with a suggested average allocation area of 2679 m^2 per 100 persons. Likewise, the average commercial land allocation could be provided on the basis of 2.29 stores and 724 m^2 per 100 persons. Moreover, he also proposed a systematic way to forecast land allocation as follows; 1) Project future population; 2) Apply land-use area ratios to the projected population and; 3) Allocate land in the city master plan accordingly (Akimoto, 2009). Similarly, Darin-Drabkin (1977) believes that land allocation should be based on rates from comparable areas, but emphasised the difficulty of predicting future needs due to different socioeconomic aspects, while future planning should be based on the assumption that less developed countries will achieve higher living standards in the future. While the ideas about comparable allocation of land have been applied widely since they were first introduced, they have been formalised into a hierarchical planning paradigm that includes the neighbourhood planning unit. But ideas about integrating population forecasting in land allocation are still undermined due to their difficulty.

Formalised planning ideas such as the Neighbourhood Unit by Clarence Perry in 1929 (Johnson, 2002) has been extensively adopted and is being used today, despite the advent of contemporary initiatives such as Smart Growth and the New Urbanism (Larice and Macdonald, 2013). The Neighbourhood Unit defines the neighbourhood's population size and borders based on the capacity of a centrally located elementary school and its walking distance from the neighbourhood’s centre (Perry, 1975). This planning unit aims to balance the supply and demand for effective operation between four functions, identified as universal functions: residential environment, elementary school, commercial shops, and small parks and playgrounds (Perry, 2013). However, the land allocation details (size and numbers) for public facilities and commercial activities were left to what is conceived as being adequate or appropriate by practitioner planners without a clear procedure to define these needs.

Clarence Perry did, however, propose some criteria for the provision of his universal functions. Most significantly, the capacity of elementary schools should be determined by modern school designs as might be acceptable by educators; at the time, he believed that an elementary school could comfortably accommodate 1600 students. Similarly, commercial land use allocations are based on one store per 100 people (Perry, 1975). The rationale for not establishing a systematic standard was because data availability was an impediment to scientific land-use planning at the time. Thus, in the instance of retail establishments, no city planner could provide an answer to the issue of how many such stores should be provided for a specific neighbourhood (Perry, 1975). Keeble (1952) took a similar approach, designing the neighbourhood unit for his New Town idea based on the location and capacity of primary schools. Keeble provided some preliminary guidance for the various functions and recognised their unavoidable inaccuracy. However, Keeble recognised the possibility of demographic instability and changes in educational policies, such as the closure of grammar schools, when it comes to school provision. As such, his ideal stance for a real-world application would be to develop a detailed education plan based on current residential accommodations, population structure, and projected growth. On the other hand, Keeble identifies the lack of precision in the provision of commercial activities as a result of remarkably little research in the field and the need for a comprehensive review of the scale of commercial provision at the neighbourhood level (Keeble, 1952). However, as the next section will reveal, future planning concepts have left these requirements unrealized.

B. Current Urban Planning

Contemporary urban planning concepts fall under different rubrics, such as ''15 minute city' (Moreno *et al.*, 2021), smart growth, sustainable development, urban renaissance (Knaap *et al.*, 2007) and the New Urbanism (Knaap and Talen, 2005). Despite the different names, they adhere to many of the same principles. These mixed use, compactness, public involvement, open spaces, walkability and transit friendliness (Grant, 2007; Knaap & Talen, 2005). When it comes to the principle of mixed use, it can be found that although the term is widely used in the planning literature, it is rarely defined (Molinaro cited in Grant, 2002). However, there are at least three conceptual levels of mixed use: 1) intensity of land use within a single category, such as planners providing various house types for different household sizes (i.e. apartments and detached houses) within one area rather than segregating them; 2) increasing the diversity of uses by having compatible land uses that support each other within an area, such as residential, education, commercial and offices; and 3) integrating some of the usually segregated uses that would be acceptable by the area users, such as garment industries near residential areas (Grant, 2002). In exploring the current practice for the methods of allocating the various uses, there is still no clear methodology that optimises land use provision. For example the concept of the SmartCode consists of seven levels of regulation settings that start with a natural zone (T1) and end with an urban core zone (T6) along with an additional special district zone (SD) (Duany, Sorlien and Wright, 2009). Figure 5 shows levels of regulation settings in SmartCode as well as levels of regulation settings in the SmartCode adopted for an illustration of the regulation settings in a typical application of the SmartCode according to the targeted development level.

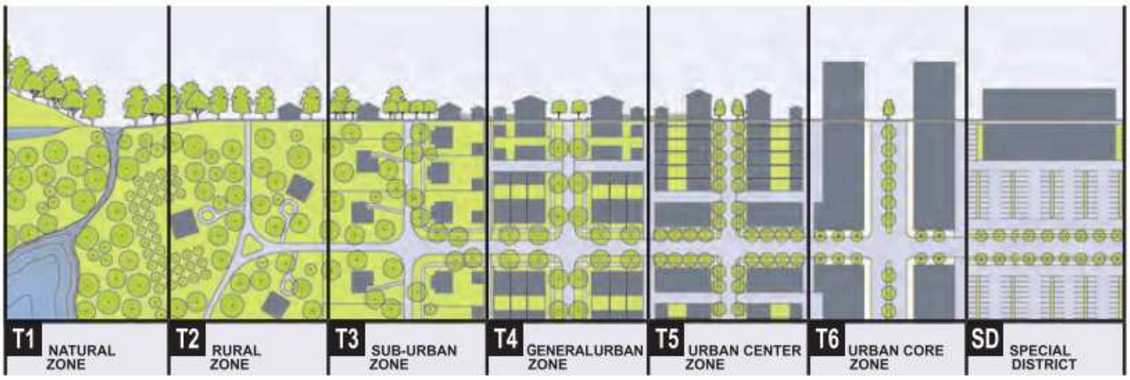


Figure 5 levels of regulation sittings in SmartCode adopted from (Duany, Sorlien and Wright, 2009).

This code's method of allocating the mix of uses is limited to a broad categorization of what would be permitted and where. Additionally, the SmartCode is intended to be calibrated to local conditions, ideally with the participation of local residents. However, the calibration procedure is not specified in the code. For example, in order to provide public buildings in zones T3 to T6, the code suggests one building lot be designated for an elementary school. It must be a minimum of three acres in size, with an additional one acre for every 100 additional housing units. Similar ready-made guides provided for commercial provision as in the general urban zone (T4), which is primarily residential, as evidenced by guidelines stating that only first-story buildings at corner locations may be utilised for retail, and each block may have just one. The specific use is limited to a 40-seat neighbourhood store or restaurant (Duany, Sorlien and Wright, 2009). It should be noted that the other land-use categories are treated similarly in the SmartCode.

Contrary to the formalisation of planning units, public participation in urban planning is becoming a more sought-after approach. However, this approach does not contribute much to the calibration of land use provision. For example, in England, more than 2,400 communities have adopted a planning scheme that has a bottom-up approach that is driven by the principle of community participation to define the uses at the local level (GOV.UK, no date). Moreover, local authorities in England use population forecasting in their preparation of local plans (Simpson, 2012). Once the local community and its municipal authority agree on the uses and spatial settings for their local plan, it becomes a binding land-use regulation (GOV.UK, 2017). Although this practice is more sensitive to the population's needs than the earlier mentioned planning concepts, the system suffers from the inexperience of district councils in forecasting the changes locally which has forced them to easily accept the developers' proposals for quality planning and demographic forecasting (Simpson, 2012). Similarly, local planning communities would lack the experience and means for future projections. This matter would strongly promote the need for a method or a tool to achieve balanced land use allocations for a long term, efficient land use consumption through compact development and reduced development costs.

C. Remarks

To conclude the urban planner's approach, the use of 'one-size-fits-all' codes repeats the same mistake of the rules of thumb practiced in the early 20th century, which is not sensitive to demographic change and the possible economic feasibility for the different functions. Moreover, the principle of mixed-use has been used as a slogan for increased diversity without a technical methodology to evaluate its outcomes (Knaap *et al.*, 2007). Moreover, public participation could contribute to defining the needs of an area, but would still require tools for balancing the demand and supply. In addition, forecasting techniques have been introduced at the local level, but it is not at the desired level of being widespread enough and nor is their quality assured. However, the economical roots for deciding the optimal location or size for the different actions such as establishing retail centres and residential zones, led to what is now widely called Spatial Interaction (SI) modelling. These have become tools for urban planners in many different planning situations for finding the best guess at an allocation in advance. There is little doubt that the urban planner needs to level up their method according to what Grant remarks; she says

> *'As professionals we have the responsibility to avoid offering simple formulas for solutions to complex problems, and to stop ignoring problems of our own creation. The task of planning the good community may not be easy: if solutions were simple society would not need professional planners' (Grant, 2010, p. 12)*

The following section will highlight the available tools that might help planners improve their approach to the issue.

2.2.2 Tools to Aid Urban Planning such as PSS

The determination of urban functions was not exclusively an urban planning problem because the solutions are more of a concern to micro-economy studies such as retail and marketing planning. These require a great understanding of human behaviour that branches into consumer and travel behaviours. However, the cornerstone element that could be considered as a key to all insights is human behaviour related to central place theory that promotes the basic behaviour of using the nearest centre assumption (Herbert and Thomas, 1982). SI theory is based on the trade-off of advantages and disadvantages to reflect the probability of behavioural decisions (Herbert and Thomas, 1982). Bid-rent theory also reflects the behavioural aspects of 'willing to pay for different locations' (Alonso, 1960). Microsimulation is based on the probabilities associated with the individuals' actions. Similarly, ABM (agent-based modelling) reflects the collective behaviour of individuals based on rule-based interactions. This is opposite to the limited ability of urban planners to plan exactly to the needs of the population while only developing arbitrary or (descriptive) land use regulations on what is to be provided and how. The economic field has advanced remarkably on this front, providing the techniques required to answer the previously asked questions related to type, size, number and establishing time. However, these answers through the SI model can only be provided in a case-by-case situation. In other words, each function (such as barber shops, laundries, schools and dentists) must be evaluated separately to decide on their economic feasibility. The techniques available from this perspective can be viewed as through the framework of PSS which can be classified into traditional and contemporary methods (Torrens, 2003) as we elaborate as follows::

2.2.2.1 Traditional methods

Traditional approaches can be differentiated by the fact that they rely on mathematical models applied to aggregated and relatively homogeneous population data. These characteristics can be seen in early SI models such as Reilly's (1929) original gravity model, Lowry's (1964) model of metropolis, Wilson's (1971) family of entropy models, and McFadden's (1975) basic logit model. However, the logit model belongs to a distinct category of discrete choice models that is more closely associated with the contemporary micro-modelling mechanisms (i.e. an individual's rational economic behaviour). The inclusion of logit models however is due to its similarity to the entropy-maximisation models. Cochrane (1975) demonstrated that the basic logit model is a special case of the single constrained entropy model. Also, Anas (1983) illustrates that the multinomial logit model is identical to the double constrained entropy model, where he states that these models have the following characteristics:

> *'In closure, "behavioral demand modeling", which follows McFadden (1973), and "entropy-maximizing modeling", which follows Wilson (1967), should be seen as two equivalent views of the same problem.' (Anas, 1983, p. 13)*

Moreover, only the Lowry model can be considered a comprehensive urban planning tool, as will be explained below. The other models are more like instruments for estimating an aggregated behaviour (i.e. origin-destination flows) for the various decisions, such as residential mobility, school enrolment, and business location. In fact, the Lowry model, utilises a gravity model for locating its household's population. These influential models, which helped pave the way for the current state-of-the-art models such as Urbansim, are crucial not only for understanding the history of modelling, but also for their continued evolution, as they continue to operate in one way or another as part of the background to contemporary models.

A. The Gravity Approach

Historically, Reilly (1929) used the gravity concept to comprehend the power of attraction of retail markets at different levels (community to metropolitan) among different towns and cities in Texas. The gravity law was expressed as follows:

> *'under normal conditions two cities draw retail trade from a smaller intermediate city or town in direct proportion to some power of the population of these two larger cities and in an inverse proportion to some power of the distance of each of the cities from the smaller intermediate city' (Reilly, 1929, p. 16)*

This study also introduces the population size thresholds that reflect on the amount of population for a function to be profitable. Although the gravitation law was not applied to calculate the thresholds in the Reilly study, later advancements made these calculations possible and robust (Wilson, 2012). The mathematical equation for Reilly's model gravitation is:

$$\frac{p_{ij}}{p_{ik}} = \left(\frac{P_j}{P_k}\right) \left(\frac{d_{ik}}{d_{ij}}\right)^2 \tag{1}$$

where p_{ij} and p_{ik} are the shares of total commerce attracted up by centres j and k from some point of consumer demand i. P_j and P_k are the populations of j and k, and d_{ij} and d_{ik} are the distances from i to the centres of j and k, respectively. One of the primary problems with Reilly's model or law is the pre-determined and deterministic method of attraction, which was later translated into a set of vector variables that measure the attractiveness power of retail centres (Barnard, 1986). Many versions of the gravitation model have been developed to overcome problems, such as the probability of a consumer choosing among overlapping trading areas (Huff, 1964) or the disaggregation of estimation by different types of population, goods, expenditure, shopping centre, and the like. Moreover, this concept has not been limited to retail attraction, but other systems of interest, such as predicting the type and size of retail markets, patient flows in health care, education enrolments, residential locations, or police resourcing for crime in small areas (Wilson, 2012). Both Huff's (Huff, 1963) and Wilson's (1971) contributions to the concept of gravity are significant in the realm of SI. As a consequence of this, there is a great deal of material that describe their construction and how they should be applied in various contexts. Whereas The Huff model for example may be utilised almost effortlessly as an ArcGIS function thanks to its direct applicability. The author is not aware of any ready-to-use entropy maximisation functions, so the coding for the model should be done ad hoc. However, Harland (2008) provides a helpful theoretical worked example that can be used as a reference to construct models of this kind.

B. The Discrete choice approach (Logit models)

Moving into the Logit model, which is substantially related to the gravity model (Bröcker, 2021), the necessity to estimate highway traffic in the 1960s prompted the development of discrete choice theories (McFadden, 1975). The core challenge of discrete choice analysis is the modelling of choice from a set of mutually exclusive and collectively exhaustive possibilities (Ben-Akiva and Lerman, 1985). The evolution of logit models lead the initial unimodal model (in terms of transport models) to become multimodal, considering price, policy, and construction (Ben-Akiva and Lerman, 1985). Also, the models have gone from being linear closed forms to nonlinear (Train, 2009). There are a number of prerequisites to working with discrete choice models, including describing the underlying assumptions, selecting the appropriate decision functions, determining parameter estimates/calibrations, and then evaluating results. The main assumption should include the following: 1) the decision makers; 2) the set of all alternatives; 3) the characteristics of the decision makers and the alternatives; and 4) the criterion for making a decision is the rule that dictates selecting an option based on its usefulness, be it monetary (i.e. net profit) or personal (i.e. comfort level). Because not all decision models are capable of addressing all of these factors, these assumptions create the foundation for selecting the proper decision model (Ben-Akiva and Bierlaire, 2012).

The models of the logit model domain can be (binary logit, multinomial logit, and mixed logit) (Brownstone, 2001) The mixed logit model can estimate any random utility and is state-of-the-art. Therefore, mixed logit models can be used for simulation modelling (Train, 2009). The advantages of using standard Logit models as well as the disadvantages of using them are as follows: 1) they permit systematic preference variance among the decision makers. However, they cannot be randomised for the various options; 2) their substitution patterns are proportionate across options with an acceptable level of accuracy; and 3) they are able to represent decisions independent of unobserved factors throughout time. Nevertheless, the correlated unknown factors cannot be resolved (Train, 2009). In response to these limitations, employing the mixed logit model will enable the following: 1) random variation; 2) correct substitution patterns; and 3) correlated unobserved factors across time (Train, 2009). The standard closed expression for the logit model is:

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{ni}}} \tag{2}$$

where P_{ni} is the probability of the alternative n in location i being selected and V_{ni} is the utility of the alternative n at i, which will be detail below. Several weighted density functions, unified or variable random coefficients, and the addition of error components can be used to transform the standard logit model into either a multinomial or mixed model with random coefficients. The mixed logit is expressed as:

$$P_{ni} = \frac{e^{\beta_n' V_{ni}}}{\sum_j e^{\beta_n' V_{ni}}} \tag{3}$$

where P_{ni} is a mixed logit probability and β_n' are vector coefficients for utility variables V_{ni}. To construct the appropriate decision model, the utility V_{ni} must be predefined based on the key assumptions and then calculated. The utility function, which describes the costs and benefits of the various alternatives of choices, is calculated using the following:

$$V_{ni} = \sum_{k=1}^{\infty} \beta_k C_{kni} \tag{4}$$

That can be extended into:

$$V_{ni} = (\beta_1 C_{1ni} + \beta_2 C_{2ni} + \beta_3 C_{3ni} + \beta_4 C_{4ni} + \ldots \beta_k C_{kni}) \tag{5}$$

Where V_{ni} is the total alternative utility from the various costs, β_k are the parameters weighting with the various costs, and C_{kni} represents the many potential costs such as price and travel time.

The calibration and validation of parameters is a significant element in the prediction of an individual's choice. This step is contingent upon the decision-makers knowledge and experience of the potential values. Consequently, the results could be deceiving if the parameters were not calibrated properly, resulting in unreasonable future forecasts. Calibration of parameters could be done by testing its estimation of choice results with a data sample (Cascetta, 2009). Several calibration techniques for the parameter β_k are available (Cascetta, 2009), yet these approaches may be avoided by using a generic optimising tool such as OptQuest (Laguna, 1997).

The prediction of changes in the future using the Logit model would consist of testing several scenarios once the primary assumptions have been defined, the utility has been calculated, the decision model of choice has been selected, and the parameters have been calibrated. Changing the assumed characteristics will result in a change in the characteristics' total utility, and the outcomes of the decision model will allow for the prediction of the probable change in choice among the set of choice alternatives.

C. A comprehensive example of the traditional models (Lowry model)

On the side of comprehensive PSSs, Lowry's (1964) pioneering Model of Metropolis, is a good example of a semi homogenous and aggregated model. Lowry's model was the first to attempt integrating a land-use transportation feedback cycle with the utilisation of a residential location model and an employment location model (Wegener, 2021). The disaggregation prospect in the model can be found in how it subdivides its key components such employment into basic and non-basic, also the land into different sectors. In addition, it applies land area constraints as an important feature in modelling (Wilson, 2016). The model's total land-use area for a specific zone is formulated as follow:

$$A_j = A_j^U + A_j^B + A_j^R + A_j^H \quad (6)$$

where A_j is the amount of land in zone j and, the superscripts represents the different sectors as U, B, R *and* H represent unusable, basic, retail and household respectively. Furthermore, the model was modular in the sense that it would take exogenous inputs from different sectors. This feature was fundamental in the model construction, as the basic employment E_j^B, which stands for the employment from exporting/industrial establishments, was predetermined and given based on the size of the basic land area A_j^B in the targeted zone j. The household population N of zone j (written as N_j) is a function of total accessible employment from both basic and retail employment. This relation in Lowry's household population model was formalised as follow:

$$N_j = g \sum_{i=1}^{n} \frac{E_j}{T_{ij}} \tag{7}$$

where N_j is the number of households in zone j, and g is a scaling factor. E_j is total employment in zone j and T_{ij} is an accessibility variable based on an index of trip distribution from zone i to j. At this point It must be recalled that Lowry (1964) theoretically compared the gravity model (laws of mass behaviour) to the market model (rational individual behaviour) to model the choice of location. Although he conveyed a strong case to use the individual behavioural approach he recognised its difficulty and higher operational cost. Besides that, there was coincidentally, the availability of data suitable to fit a gravity type model. In fact, Lowry believed that if the laws of mass behaviour are dependable, then they must reflect the aggregation of individuals behaviour and he stated:

> *"The traditions are not necessarily antithetical in content, although they certainly are in style. If there are dependable "laws" of mass behaviour, they certainly must represent the aggregate outcomes of myriad individual acts; given time, and saved from logical errors, the location theorists ought to come out at the same place as the social physicists" (Lowry, 1964, p. 22)*

The model consisted of 12 equations that were iteratively programmed to create an instant metropolis (Lowry 1964). The model's static equilibrium is a property of early models for dealing with time. Moreover, they influenced many of the present more complex land-use transportation models (Wegener 2021). According to Iacono, Levinson and El-Geneidy (2008), some of these models are founded around the Lowry model, including TOMM, PLUM, ITLUP, LILT, and IRPUD. However, for a thorough grasp of the model's composition and workings, see Lowry (1964), as well as for the process of refining the model into a more contemporary state of art practice, see (Wilson, 2016). Finally, for a review of recent comprehensive models see (Acheampong and Silva, 2015; Hunt, Kriger and Miller, 2005; Wegener and Fuerst, 2004). The next section will discuss more contemporary approaches in urban modelling.

2.2.2.2 Contemporary Methods

The second wave of urban modelling departs from the often aggregated (top-down) mathematical models such as the SI models into disaggregated (bottom-up) programming models like ABM which forecast temporal changes (Torrens, 2003). According to Heppenstall *et al.* (2012) the ideas of considering the actions of individual actors including geographical representations along with the flattered progress of homogenous aggregated models made the concept of agents a mainstream research field. This departure from traditional modelling has the advantage of considering: 1) the real behavioural dynamics of individuals and entities; 2) the interactions between the system actors; 3) the dynamic feedback within the system's actors; 4) scalability of modelling levels; and 5) greater flexibility in model building (Torrens, 2003). In the area of complex computer simulation models, cellular automata, ABM and microsimulation are distinct methods but with common ground (Birkin and Wu, 2012) is that these approaches are based on individuals' interactions. In this part, the emphasis will be placed on ABM and microsimulation as the primary methods utilised in the process of developing the research models and as a result of their growing adoption in recent years. Whereas microsimulation is primarily driven by statistics and probability, agent-based models are mainly rule-based focused (Ballas, Broomhead and Jones, 2019).

In a manner analogous to that of the conventional approaches, where there are operational mechanisms such as the Gravity or Logit models and compressive models that use those processes inside an urban modelling framework such as the Lowry model, we could consider the following: Both ABM and Microsimulation may be thought of as mechanisms that can be used to generate comprehensive models like the UrbanSim model by Waddell (2002), and one of the leading examples of these models is UrbanSim in terms of using discrete choice theory, fine levels of scale, and an agent-based focus (Batty, 2011). To name other examples, there are: the Integrated Land-Use Modelling and Transportation System Simulation (ILUMASS) (Moekel *et al.*, 2003) and the Integrated Land Use, Transportation and Environment (ILUTE) (Salvini and Miller, 2005) are related state of the art models which are related to state of the art models.

A. Agent-Based Modelling

This section will provide a complete overview of the ABM approach, focusing on its fundamentals from an urban modelling perspective. In addition, where appropriate, diversions are made to cover relevant issues, such as the sub-models of population growth and residential mobility or the description of the optimisation algorithm utilised inside the employed ABM platform. Furthermore, the author have observed, that Microsimulation models are frequently utilised in the specific area of population estimates as an implicit part of ABM models or independently modelled, and more particularly for the synthesis of small-area population. As a result, this method will be discussed as part of the ABM approach within its key component of base year population as to how it has been utilised in the process of developing this research model.

A.1 History

In the late 1950s, Guy Orcutt introduced microsimulation to the social sciences by developing a "new type of economic system." (Arentze, Timmermans and Veldhuisen, 2010; Birkin and Wu, 2012; Wegener, 2021). Meanwhile the origins of agent-based models can be traced back to 1970s cellular automata models, such as Schelling's (1971) distinguished attempt to simulate human and social behaviour. Over time, ABM gained widespread recognition, if not widespread application, in the 1990s, and is now acknowledged as a standard modelling methodology across a wide variety of fields (Manson *et al.*, 2020). It also can be noted that the complexity of ABM models has gradually increased over the years (Ballas, Broomhead and Jones, 2019). According to Crooks, Patel and Wise (2014) recently lead academics in urban systems are also looking at the outcomes of the decisions of individual agents to comprehend urban planning issues like congestion and sprawl. Moreover, technological and conceptual advancements, have made the fusion of agent-based and microsimulation in urban models possible and these have spread globally (Wegener, 2021). According to Wilson (2012) the outcomes using an agent-based model can be plausible with respect to more aggregate mathematical models and it is worth examining further the equivalent outcomes at fine scales of these different modelling approaches. On the other hand, the same may be said regarding the popularity of Microsimulation. The author have observed, however, that Microsimulation models are frequently utilised in the specific area of population estimates as an implicit part of ABM models or independently modelled, and more particularly for the synthesis of small-area populations. As a result, this method will be discussed as part of the ABM approach within its key component of the base year population as to how it has been utilised in the process of developing models that underpin this research.

A.2 Definition

ABM can be defined broadly as a computer representation of an actual system (Auchincloss and Diez Roux, 2008) with the notion that any moving or static agents in the system acts in accordance to their behavioural profile (Araújo de Oliveira, 2022). However, the definition of ABM may be expanded to incorporate more defining characteristics, such as the capacity of agents to communicate, interact, and reproduce, as well as the possession of skills and resources (Gilbert and Doran, 2018). In addition, being behaviourally adaptive to its surroundings (Ben-Dor, Ben-Elia and Benenson, 2021) in a manner that is not centrally controlled (Crooks and Heppenstall, 2012; Torrens and McDaniel, 2013).

A.3 Importance

Due to exorbitant expense, safety concerns, or other factors, it may not be practical to test a potential solution to a real-world situation using mathematical models (Borshchev, 2013). Consequently, one of the approaches for assessing potential solutions should include simulating the actual world in a virtual environment (Railsback and Grimm, 2019). The general objective of ABM is to examine the dynamics of diverse agents statistically and qualitatively (Ben-Dor, Ben-Elia and Benenson, 2021). Besides these important features, perhaps one of the key advantages of ABM is the trackability of agent characteristics of interest, while simulations run and interactions occur (Gorman *et al.*, 2006) while this also allows the investigation of change at a small area level (Ballas, Broomhead and Jones, 2019). This feature goes beyond descriptive analysis and into hypothesis testing to provide useful insights into the topic at hand (Cerdá *et al.*, 2014). Another important feature is the flexibility in model structuring and specification to target specific problems in its context (Batty, 2012). However, this feature makes it rare to find highly standardised models for similar problems (ibid). This matter brings us to the importance of understanding the main modelling components and specifications which will be discussed below.

A.4 The models specification

Modellers must incorporate all necessary details in their models even if they are a simplified representation of reality. This ensures the accuracy of their output (Borshchev, 2013). This can only be determined by doing a series of experiments and comparing the results to the expected outcomes (Railsback and Grimm, 2019). ABM models are comprised of elements such as base year population, types of space, types of time, scale, updates, model intentions, included sub models, utilised decision functions and visualisation methods. See Manson *et al.* (2020) for a comprehensive review of main methodological challenges in spatial agent-based modelling. However, in this part the focus will be on some of these components according to their relative importance to the models which have been constructed models in this thesis.

A.4.a Base-year population (Microsimulation)

According to Heppenstall and Smith (2021) selecting base year population is regarded as one of the primary challenges because of the usual lack of population data. Whereas the available data frequently include the population's overall statistics (i.e. total counts of males and females), the co-dependencies of the population's characteristics (i.e. household members, income and education level) are rarely available, particularly at the small area level. Moreover, what is often accessible is a sample of the population that contains all the sought-after characteristics. Given the usual lack of data, the role of microsimulation becomes evident as a means of synthesising the necessary population data. It is possible to employ microsimulation to give ABM's agents their initial attributes and positions with this method. So, with the help of ABM, individuals' adaptive behaviours and the emergence of new behaviours might be modelled (Briassoulis, Kavroudakis and Soulakellis, 2019). Microsimulation allows data from different sources to be integrated and trends to be investigated at different geographical scales. In addition, they permit updating and projection, which is crucial for predicting future population trends (Ballas and Clarke, 2001). As stated in Heppenstall and Smith the definition of Microsimulation is:

> *"Microsimulation is the generation at time* $t = 0$ *of a population sample* P *made up of* n *individuals* $[p^1, p^2, \dots p^n]$ *where each individual,* i*, has a number of initial attributes* $[a_1^i, a_2^i, \dots, a_m^i]$*. The population is then updated to later times,* t*, so the attributes of individual* i *become functions of time* $[a_1^i(t), a_2^i(t), \dots, a_m^i(t)]$ *." (Heppenstall and Smith, 2021, p. 1769)*

In this sense the population is dynamically changing, but not assigned with location nor conforming to the characteristic observed in another data set for creating a population synthesis. To include these features in the mathematical definition Heppenstall and Smith define the following:

> *"Mathematically, the population of area* j *can be written as* $P_j = [w_{ij}\ P^i]$ *where the weight* w_{ij} *represents the number of people in the population of area* j*,characteristic* i*, with the attributes of person* P^i *from the sample or synthetically generated population." (Heppenstall and Smith, 2021, p. 1769)*

According to Harland *et al.;* Heppenstall and Smith (2012; 2021), the most commonly used microsimulation algorithms for population synthesis are Deterministic Reweighting (DR) (Ballas, D., Rossiter, D., Thomas, B., Clarke, G. and Dorling, D., 2005) Conditional Probability (CP) (Birkin and Clarke, 1988) and Simulated Annealing (SA) (Openshaw, 1995). Regarding the DR method, which relies on Iterative Proportional Fitting: (IPF), this method iteratively fits each record in the sample population in proportion to the observed population counts of the small area until all constraints are taken into account. This fitting process begins with demographics that are generally equally divided throughout categories and progressing to those that are less evenly distributed. As the name suggests, the method does not utilise stochastic approaches. But the sequence of attributes fitting would control its outcomes (Heppenstall and Smith, 2021). The CP method generates a population where joint probabilities of associated characteristics are derived from constraint tables. Individuals' attributes are assigned sequentially, similar to the DR approach in terms of the demographic's distribution. This method was initially created to construct a synthetic population without survey data (Harland *et al.*, 2012). The SA approach creates a population as a random extract from sample population data, and by aggregating for constraints, the population's fit to constraining tables can be assessed. Then, a random member of this population is substituted with another random member from the sample. If the fit improves, the new member replaces the previous one (Harland *et al.*, 2012). Contrary to the DR algorithm, both CP and SA are stochastic models.

If implementing a microsimulation method is required for operational models, there are some aspects to be considered for the generation of population data. In a sense these aspects are not exclusive to microsimulation but, strongly related in this context which includes; 1) the best initial data for such a process would be collected particularly for modelling purposes with the proper scope and amount of spatial disaggregation (O'Donoghue, Morrissey and Lennon, 2013); 2) the tremendous data requirements and significant processing time make large-scale implementation of these models difficult (Moeckel *et al.*, 2018); 3) The stochastic base of these models causes lack of stability and as a source of error that increases with large choice variation and less evident preferences of agents, therefore it may offer a misleading feeling of correctness (Wegener, 2021); and 4) the static or dynamic nature of the created population, while the former produces population with their characteristics. Based on the interactions between the individuals, the latter cause future changes (Heppenstall and Smith, 2021). Other aspects need to be considered for the ABM models and these follows.

A.4.b Types of space

Conceptualization and representation are two of the most essential components of space (Manson *et al.*, 2020). There are two fundamental ways to conceptualise space: absolute and relative. While an absolute space is a zone with distinctive attributes (which is utilised in the form of mass attraction in gravity models), relative space represents the relationship between items in the absence of defined zones for location attributes (which is used in the bid-rent theory or the general distance decay function). On the other hand, space can be represented either implicitly or explicitly. In contrast, implicit space models are non-geographical in the sense that they emphasise network dynamics and are only partially concerned with location; for example, agents with group affiliation are more likely to exchange information or a network of flows that utilises cost and distance with geography being implicit. The explicit models, however, are geographical. With the use of concepts such as two-dimensional grids, explicit models can be an abstraction of reality. In contrast, a realistic representation employs vector mapping to reflect the actual space geometry. However, space representation is complicated by the need to choose between discrete and continuous measurements that would be based on the theoretical modelling approach and the available data.

A.4.c Types of time and updating

The advancement of the model is dependent on temporal modelling and the update of agents. The time types will be covered first, followed by the agent updating types.

Most ABMs use equal-time or event-based time modelling (O'Sullivan and Perry, 2013). The equal-time technique maps each model iteration in equal-duration time steps. Event-based simulations determine the time between events in a queue. Recursive models, in which the end state of one iteration serves as the starting point for the next, are the most common models (Wegener, 2021). This is due to its conceptual simplicity since all modelling processes work at the same pace (Manson *et al.*, 2020). In recursive models, many equilibrium time points are constructed prior to the target year. In which the modifications from the integrated sub-models may have varying equilibrium speeds (Wegener, 2021). According to Hutchinson and Batty (1986), the categorization of equilibrium speeds in urban models includes: 1) Slow Processes, which may be represented by building projects that have a lengthy lifespan and take many years to complete; 2) Medium-Speed Processes, such as ageing and birth; and 3) Fast Processes, which occur daily or will be completed in less than one year.

O'Sullivan and Perry (2013) divided updates into three types: the first is perfect synchronisation, which is realistic and computationally expensive since it forces the modeller to make numerous judgments about how to simulate a process when two agents access one resource simultaneously. The second technique, random asynchronous updates, requires agents to update their states and act at each time step in a random sequence. Randomness is utilised when principled or empirical ways to ordering agent behaviours are lacking. The third technique, ordered asynchronous updates, is preferred when the precise sequence of events can be known by theory or empirical observation. Changing rules or adding new environmental conditions might be timed.

A.4.d Intention

The goal of developing an ABM might be either predictive or explanatory. The objective of predictive models is to build a representation of the underlying environment that is as accurate as possible, so that the results may be empirically verified. On the other hand, explanatory modelling focuses on enhancing the theoretical explanations of a phenomenon (Macy and Willer, 2002). In addition to developing predictive capabilities, ABM may also help us get a better understanding of basic processes (Axelrod, 1997).

A.4.e Included Sub-models

Given the ability of ABMs to integrate different sub-models that form a complex system such as the urban system, it would be appropriate to touch upon the level of abstraction and the sub-systems that could be considered while constructing a comprehensive urban system.

According to Manson *et al.* (2020) in order to achieve an 'acceptable' level of model complexity, most modellers believe that the complexity of the target system must be taken into account while developing the model. However, there is no consensus on how to get to an "acceptable" level of complexity. Typically, there are two approaches to handle this issue. First, models should be made as simple as is practically feasible, and complexity should only be introduced if the model is unable to correctly describe the system in its most basic form. Second, models should be developed in a sophisticated way to represent evidence and knowledge about the target system, which then gradually reduces unnecessary components (Manson *et al.*, 2020). Another aspect to consider in the complexity of systems is the different genres of sub-models as their agents may be built to have a particular generative behaviour (i.e. rule based actions) or fitted estimates of observed data (i.e. statistical estimates) (Parker *et al.*, 2003) which is subject to data availability and the knowledge level about the modelled system.

Process graphs are commonly used to demonstrate models flow work and the links between the sub-models in a system. However, these graphs represent the abstraction of the targeted system given the model's limitations and not the complete working process of the actual system. It is a challenging task if not impossible to create a process graph of all sub-models that would make a comprehensive urban model. This is because of the complexity of interconnections between the various sub models, The causation and correlation dilemma of barely understood dynamics along with the multi-pathway processes and the conceptualisation of hierarchal sub models inside the principal sub-models, despite the fact that our capacity to comprehend the complexities of the urban systems, is restricted. However, it is necessary to have a comprehensive and in-depth understanding of the system that is being targeted as well as the sub-models that comprise it. The sub-models that might be found in the literature of urban modelling are listed here. Based on the reviews of Hunt, Kriger and Miller (2005) and Wegener (2004), this list was created with the intention of providing guidance on the primary sub-modes and some of the aspects expected within each rather than being complete or inclusive.

1) Land-use polices modelling which cover all general laws and regulations such as taxation, environmental sustainability, building regulations, land zoning and land consumption,
2) Economic modelling at both global and local scales which includes major industries also, local shopping as services provided, and employment pools,
3) Demographic modelling that is concerned with population growth, social change and population profiling,
4) Transportation modelling that is sensitive to the accessibility levels of the infrastructure and aims to reproduce the daily trips,

5) The modelling of the housing market which includes developers' activity, the prices, the housing conditions, and the housing and land availability,
6) The modelling of residential mobility that is associated with the life cycle of individuals, their established social capacity and income, along with the specifics of the next move triggers, and the choice of next move location.

As mentioned before, these sub-models are strongly interrelated. For example, a location with limited accessibility may see a slower rate of population growth, which is projected to reduce the need for additional housing units. Due to the limited demand, developers will lose interest, and land usage will be minimal. In reality, this feature is very narrow-minded given the number of factors that are altering the city's dynamics. Moreover, each of these sub-models is regarded as a separate research field; hence, the aims of urban models (i.e. planning support systems) are typically centred on particular challenges. As a result, the inclusion of sub-models is logically restricted in order to reduce complexity, and simpler statistical projection methods can replace more sophisticated models.

This study centred on modelling population demography and residential mobility as the primary determinants for estimating the number of students at both the city and district levels. In addition, land-use policy and the housing market were modelled as restrictions. These constraint models were projected statistically for simplicity. In addition, economic and transportation modelling were overlooked for the current version of the developed model. In the following sections, the included sub-models will be discussed in detail.

On the population side of sub-modelling, population growth and mobility were further studied as below:

- Population growth modelling

On Population growth estimation*,* A wide range of methods available for population projection are dependent on extrapolation and interpolation methods. However, no single model is suitable for all purposes and needs. Therefore, the choice of the right model is subject to aspects such as data requirements, cost and ease of application. For example, any population estimates that require detailed age data will need to adopt a cohort approach (Swanson and Tayman, 2012). Virtually CCM is the most used approach everywhere (Rogers, 1995) and, generally recommended as a standard method by the United Nations (Simpson, 2017). According to Booth (2006), the standard CCM is a demographic accounting system that calculates population changes based on rates of survivorship, births and migration. Therefore, these rates split the population projection process into three forecasting tasks (Lee, Carter and Tuljapurkar, 1995). Moreover, this method is usually accompanied with three deterministic scenarios as high, medium and low, resulting from combining the factors of change (Booth, 2006). However, the deterministic approach can be entirely dependent on a single component (usually fertility). Current application of CCM are pushing towards more probabilistic than deterministic approaches (Ševčíková *et al.*, 2016). CCM are known to be data demanding and computationally intensive (Swanson 2011). The limitations of this approach lies in the use of fixed correlated age groups, the use of fixed range of probability, the inability to specify the probability for the high- low intervals (Booth, 2006). Given the uncertainty level surrounding the demographic change and the rigidity of the scenario-based approach CCM outcomes should be conceived as a projection instead of a forecast (Booth, 2006). Swanson and Tayman (2012) Identifies five types of component models However, the fundamental demographic structural equation (standard CCM) can be written as follows:

$$P_{i,t+k} = P_{i,t} + B_i - D_i + I_i - O_i \tag{8}$$

Where:

$P_{i,t}$ Population of area i at time t (base year)

$P_{i,t+k}$ Population of area i at time $t + k$ (the estimated year)

B_i Births in area i between time t and $t + k$

D_i Deaths in area i between time t and $t + k$

I_i In-migrants in area i between time t and $t + k$

O_i Out-migration in area i between time t and $t + k$

Swanson and Tayman (2012) provide a step-by-step explanation of the method.

In forecasting the components of population change namely fertility, mortality, and migration, the main approaches can be split into three categories as stated by Booth:

> *"Extrapolative methods focus on the regularity of patterns and trends and extend these into the future without recourse to other knowledge in the form of exogenous variables. Methods based on expectation may use individual data (such as surveys of women's future birth expectations) or the opinions of experts about future demographic developments. Methods seeking to explain demographic processes use structural models based on theories relating demographic quantities to other variables." (Booth, 2006, p. 550)*

These projection approaches can be used separately or combined (Booth, 2006). However, deciding on the best model to use is challenging for any of these components

(Bohk-Ewald, Li and Myrskylä, 2018; Booth, 2006; Shang and Booth, 2020). This is generally because the data can be of non-stationary nature (Booth, 2006; Keyfitz, 1972; Swanson and Tayman, 2012), and the accuracy of the different methods are case specific (Booth, 2006). Thus, the method's accuracy are not always comparable (Schoen, 2016), the estimation of uncertainty remains questionable (Booth, 2006), and there is unpredictability of turning events (John Bongaarts and Rodolfo A. Bulatao *et al.*, 2000; Rogers, 1995). An example for sudden changes can be usually found in migration data which is subject to very sudden turning points due to political and economic changes (Vanella, Deschermeier and Wilke, 2020). Another example for the non-stationary data can be found in period fertility (TFR) data, which is known to suffer from temporal fluctuations (Booth, 2006; Schoen, 2004). However, it seems that extrapolation methods via the Autoregressive Integrated Moving Average (ARIMA) model and its variants are very key for projecting the components of population change (Booth, 2006; Shang and Booth, 2020; Swanson and Tayman, 2012). For example, the role of the Lee Carter method which is special case of the ARIMA model (Hyndman and Shahid Ullah, 2007) has been considered as the most important method for projecting mortality rates in the US and widely adopted globally (Girosi and King, 2007). Likewise, but with slightly less momentum, the Lee Carter model has been extended and used in fertility projections. See Girosi and King (2007) for a detailed explanation of the Lee-Carter method. In this research, we focused on the Lee Carter variants for they continue the progress to overcome the original method's shortcomings and there is convenience in using them in the R programming environment.

On one side, according to Swanson and Tayman (2012), extrapolative time series models are challenging to apply, should be handled by high level statical experts and require relatively long historical observations. There is no minimum number of observations but, they have been argued to range between 30 – 60 observations (Swanson and Tayman, 2012). For the original Lee-Carter method, the main critique was that it produced unreasonable forecasts for particular age groups (Wiśniowski *et al.*, 2015) such as age-group 80 and above (Booth, 2006). Another downside is that extrapolative methods are not meant to model structural changes but to generate a large spectrum of potentials from past change (Booth, 2006). Therefore, it should not be expected to achieve a structural level of predictability. On the other hand, this type of model only requires historical observations and they facilitate the provision of confidence levels around its estimates (Swanson and Tayman, 2012). According to Schoen (2016), the differences among the

Lee-Carter variants belong to two categories, the treatment of input data and the specification of time series modelling. In the category of input data, the change can be on how the data is processed as being raw, in logarithms, and/or smoothed, with the choice of parameters to adjust fitted data and the choice of transition smoothness from present to forecast data (the jump off point). On the time series specification side, the original model uses a random walk with a drift ARIMA model that utilizes a single principal component. However, it can be advanced by incorporating different statistical ideas into the ARIMA model such as using multiple orders of principal components instead of first order principal components. The Hyndman and Shahid Ullah (2007) model is an example based on an extensively modified structure that uses function data nonparametric smoothing and robust statistics.

- Residential mobility modelling

On the other side, residential mobility is the study of the economic decision to relocate and, if populations do, they make new housing choices (Brasington, 2021). This area of research is crucial for urban modelling because it can provide answers to hypothetical issues that are based on population demography (Arentze, Timmermans and Veldhuisen, 2010). The three basic approaches for modelling residential mobility are life-cycle, cost-benefit and, neighbourhood change (Winstanley, Thorns and Perkins, 2002).

According to Coulombel (2011), in life-cycle models, common life events such as seeking higher education, getting married, having children, and career changes are central into considering home moves. In fact, it has been observed that patterns of movement are generally consistent in relation to life events and age. For instance, young individuals between the ages of 20 and 35 are the most mobile population cohort. The cost-benefit kind of model, on the other hand, is an econometric technique that considers all potential costs and benefits involved with relocation decisions (John, Dowding and Biggs, 1995). This involves the value of the existing location as it is, the costs of improving the current location, and the costs and benefits of moving. In addition, this approach is dependent on a "threshold of dissatisfaction" that motivates those who are willing to relocate to examine alternatives while considering their existing location as a basis for comparison (Shumaker and Stokols, 1982). For the neighbourhood change models, surrounding elements and the interrelationships between the decision maker and their communities are the focal points. This is due to the fact that many individuals are reluctant to abandon convenient and familiar surroundings to which they have accustomed to and established strong connections (Winstanley, Thorns and Perkins, 2002).

However, these models are not perfect in the sense that the intricacy of residential mobility has been oversimplified in order to satisfy some statistical analysis requirements (Arentze, Timmermans and Veldhuisen, 2010). This is also because none of the factors that determine movership can be considered independent of the others since they are all interconnected and interdependent (Jordan, Birkin and Evans, 2012). Likewise, these models are affected by the use of 'static' secondary data that fails to capture the longitudinal moves of residential mobility. Moreover, research focusing on individual persons or individual households do not provide information about all household members (Winstanley, Thorns and Perkins, 2002).

Despite the shortcomings of these models, some factors can provide valuable predictability. The family life cycle (Dieleman, 2001) and income appear to be the most important elements in influencing these processes (Boehm, 1982; Dieleman, 2001). In addition, there is a direct association between the amount of disposable income that can be spent for housing and the cost of the home (Jordan, Birkin and Evans, 2012). Typically, three decisions are predicted in residential mobility modelling: 1) whether to move; 2) where to move; and 3) what to move into (Brasington, 2021). In addition, these decisions can be modelled sequentially, as in a layered framework (Fischer and Aufhauser, 1988), or concurrently, as in a series of simultaneous decisions (Brasington, 2021).

According to Brasington (2021), gravity or discrete choice models could be used to simulate residential mobility, with the latter method being the most used at present. The discrete choice framework is used to simulate the related decisions such move-or-stay decisions, the rent-or-own decision, and the selection of a new residence among a variety of places. The discrete choice framework is also utilised in more complex procedures, such as sequential choices of move versus stay, followed by selection of a specific neighbourhood from a finite set of options, and then selection of a specific residence inside the neighbourhood (Brasington, 2021). In particular, a move is triggered when the disparity between the existing and ideal residences surpasses a stress threshold (Coulombel, 2011). According to Jordan, Birkin and Evans (2012), depending on the complexity of the life-cycle model to be employed, one may choose to model the complete family as one agent (i.e. a household) or model every potential person while utilising the unit household, to represent an aggregate of persons. In addition, on one side, the decision to relocate individuals is modelled based on demographic factors such as age, income, race, and household size. On the other side, for residences choices, we can utilise characteristics such as accommodation type, value, vacancy status, etc. See Fotheringham *et al.* (2004) for an illustration of the intricacy of modelling a similar problem such as migration.

A.4.f Decision functions

ABM typically uses one of the following three methods to make decisions (Kennedy, 2012): the first is based on mathematical models, whereas the second is founded on a cognitive framework. Both these methods can be viewed as rule-based. The third method, based on cognitive architecture models, which depends on a bottom-up artificial intelligence approach.

The mathematical approaches generate agent behaviour by employing mathematical simplifications and threshold-based rules to determine outcomes (Kennedy, 2012). On the basis of benefit maximisation and cost minimization (Coleman, 1994), the results would be understandable and comparable to human behaviour (Kennedy, 2012). An example of such models could be the Logit models which were discussed previously. Taking into account the definition that is employed here to present the mathematical models, the author argues that certain forms of Machine Learning (ML), such as Neural Networks (NN), could be included in this classification. As a result of the similarity between the method's simplification of brain mechanisms, this is achieved through the calculation of connected weights, and the use of threshold activators in order to make decisions with the highest probability of occurring. It is essential to emphasise, however, that this basic mechanism of abstraction and thresholds can be found in the other two approaches as well. The dividing line between these approaches is their theoretical foundations. Conceptual frameworks are based on the hierarchical modelling principle (Schmidt, 2000; Singh, Padgham and Logan, 2016) whereas Cognitive Architecture is founded on creating an autonomous system using ML methodologies that are behaviourally adaptable to their environment (Panella, Fragonara and Tsourdos, 2021).

At this point, it would be appropriate to introduce the employed ML algorithm as a behavioural decision-making technique with potential application in urban modelling, given the type of available data and the discrete nature of decisions to be made in this research. The focus in this section will be on the NN algorithm for predicting multi-class classifications. Based on the initial findings from testing several machine learning algorithms, NN was chosen as the best option in terms of prediction accuracy.

Machine learning algorithms can be viewed as computing systems that utilise data to produce accurate decision or prediction (Jordan and Mitchell, 2015). Moreover, ML are used to address issues such as classification and regression. The key advantage that any machine learning model offers is its capacity to generalise its predictability from a finite set of previously seen training data to unseen instances (Aggarwal, 2018). The three primary categories of ML are Supervised Learning, Unsupervised Learning, and Reinforcement Learning (Tiwang, 2020). The classification algorithms include K-Nearest Neighbours, Decision Trees, Logistic regression, Support Vector Machines, Naive Bayes, and NN. According to Tiwang (2020), a major factor in the widespread usage of NN is the fact that NN regularly outperform alternative models in a wide range of applications.

In the context of classification problems, a binary classification problem is a classification with only two class labels. In contrast, a multi-class classification problem involves the classification of more than two classes. However, one of the most significant drawbacks of classification systems is that they can only identify what they have been trained to recognise (Krogh, 2008). For example, predicting the possibility of a household to move to a newly established district cannot be done due to the lack of its class label in the trained model. Another related issue would be the absence of training data to update the model at an early stage. However, a partial solution to this issue could be by introducing additional classes such as "other" or "unknown".

An artificial neural network comprises an input layer, one or more hidden layers, and an output layer. For illustration See Figure 6 A typical NN design for a multi-class classification problem.

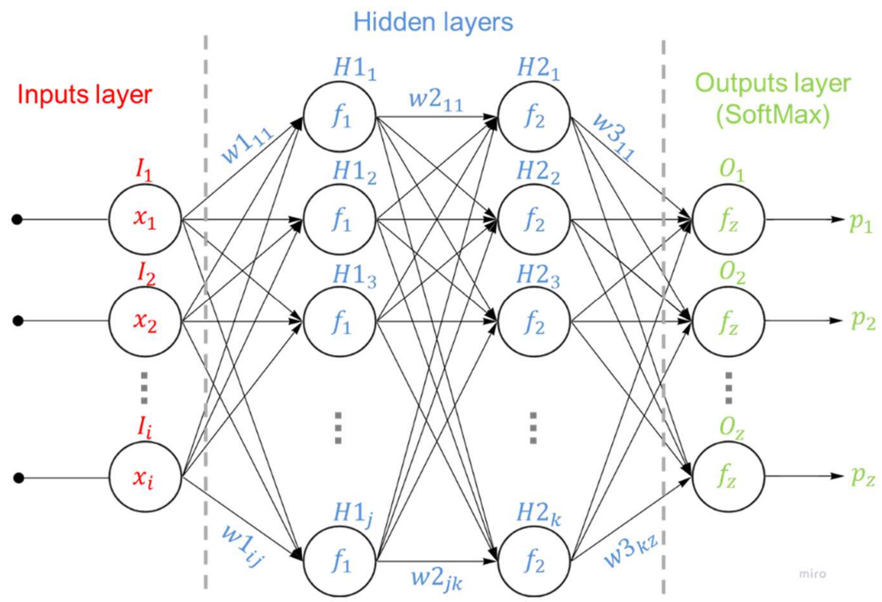


Figure 6 A typical NN design for a multi-class classification problem.

Figure 6 shows that information always travels forward in the network, starting at the input layer (represented by nodes I) and continuing through the hidden layers (represented by nodes H) to the output layer (represented by nodes O). The fundamental components of NN are nodes, also known as Neurons. The connections between these neurons are weighted links (represented by links W). When the inputs enter the neural network, the weight of each link is multiplied by these inputs before being processed by the neurons. After that, for each neuron, the inputs are summed then applied to an activation function (represented by f), which results in the output of the neuron. This process of neurons input-output is repeated as many as the number of neurons in the network. The last layer is a usually a SoftMax function (represented by f_z), which simply converts the outputs of the final hidden layer into probabilities (represented by p). The SoftMax function uses the standard logit model, which was shown in Equation 1 earlier. This aspect shows that there is a fundamental similarity between NN and logit models when it comes to turning inputs into probabilities. However, to compare these models from an operational point of view. One could argue that some of advantages of using a logit model could be: 1) the manual formulation of the model where the considered aspects are known with their affect in a positive or a negative way; 2) the weights are directly associated with the inputs, which makes it easier to interpret the weights. However, the disaggregation of population characteristic would require creating different versions of the model; 3) the calibration of the logit models is a separate and daunting process, whereas the advantage of using a NN approach could be in the automation of the learning process. Once the number of layers and nodes are defined for the NN, the modeller will have to overcome the need for formulating the relations among the inputs and their weights; and 4) the disaggregation of the population could be done easily if their characteristics are included with the inputs. However, the weights are based on a network of links that meant to be legible by a computer in the first place. Therefore, the NN approach could be viewed as a soft of black box.

Typically, In NN there are no feedback weights linking the next layer to the previous layer (Wang, 2019) or connecting the nodes within the same layer (Aggarwal, 2018). A neuron function f is basically a non-linear mathematical function known as activation function applied to the inputs it receives (Tiwang, 2020). Several activation functions are available such Rectified Linear Unit (ReLU), Hyperbolic Tangent (Tanh) and Sigmod (Aggarwal, 2018). However, ReLU's ability to avoid vanishing and exploding gradient problems has made it increasingly popular in recent years (Aggarwal, 2018). Moreover, in the hidden layers, the number of nodes and layers is normally arbitrary chosen (Thunberg and Mannerskog, 2019). On the contrary, the number nodes in the output layer are equivalent to the number of prediction classes (Tiwang, 2020). According to Ding *et al.* (2013) The performance of neural networks is known to be sensitive to the number of nodes in the hidden layers. Too few nodes might result in a poor approximation, while too many may lead to overfitting issues. Clearly, there is a trade-off between improving network performance and simplifying network topology. However, some algorithms could be used to optimise the architecture and parameters of neural networks to overcome this problem (Whitley, 2001).

The training of Supervised learning is achieved by iteratively modifying the weights between nodes over a considerable number of input-output pairs, the function produced by the neural network is tuned over time to provide more precise predictions (Aggarwal, 2018). This process includes the establishment of a cost function to reflect the error between the true class label the NN's predicted class label. This error is intended to be minimised so that the number of true classifications is as near as feasible to the actual classifications while the backpropagation algorithm is used to minimise the cost function (Tiwang, 2020). Even though several methods are established throughout the training process, such as Newton's technique, stochastic gradient descent and etc, they may all be considered gradient descent algorithms under certain circumstances (Wang, 2019).

To determine the generalisation performance (i.e. accuracy) of a NN, it must be evaluated on data that were not used to train the network. Typically, this is accomplished using cross-validation, in which the data set is divided into smaller non-overlapping sets for rotational training and testing, such that all sets are tested. This provides an estimate of the network's generalisation ability (Krogh, 2008). Overfitting is a widespread issue in NN design, resulting in networks that perform very well on training data but poorly on unseen data (Aggarwal, 2018).

Among the multiple ways to evaluate a multi-class classification model is the Top-k error measure (Lapin, Hein and Schiele, 2018). This measure ranks all prediction classes based on their likelihood of being true, then only if the correct class is among the Top-K classes, this implies a true prediction. Therefore, when the class with the highest probability is considered, the top-1 error is calculated. Therefore, Top-k error can be employed when the targeted class is regarded to be among a short list of k classes sorted by likelihood. The author believes that this type of error metric, could be employed when decisions are nearly deterministic in nature. For example, choosing a specific location for the next move in a residential move may not always be achievable due to rent costs or unit availability in that area. Instead, a short list of areas is typically employed, with a degree of equal or non-equal chance influencing the final location. This method prevents low prediction accuracy from being concluded for certain issues when the truth is not deterministic (i.e. in one option). In cases when the reality it is not deterministic, this method keeps us from jumping to conclusions about our forecast accuracy (i.e. considering the option with highest probability). Shortlisting the top five or top three candidates from a large pool of options is an excellent place to start. Now, let us hope that the right answer is among the shortlist, and that the other four were actually considered by the decision maker.

Scikit-learn, TensorFlow, and DeepLearning4j are just a few accessible ML tools that can be employed (Aggarwal, 2018). For Python users, Scikit is an easy-to-learn ML package. Additionally, TensorFlow is an advanced open-source platform option, while on the other hand, DeepLearning4j is a Java-based programming library.

A.4.g Visualisation

One of the significant advantages of using ABM is its visual capacity over the traditional methods which normally have their outcomes presented as charts and choropleth maps. The visualisation of an ABM could be extensively detailed to cover aspects such as the schedule of events as what might be an interacting flowchart, the agents as mobile or immobile objects, the physical and nonphysical links between the agents and the model outcomes as transitional changes in the appearance of objects and their locations. All these possible visualisations provide more insight beyond what can be offered by a typical visualization of charts. The benefits of such advanced visualisation capacity may enhance the model transparency (Batty, 2007), conveying and comprehending the modelled phenomenon's relevant behaviour (Kornhauser, Wilensky and Rand, 2009), enable the visual spotting of errors to verify and validate the model outputs (Crooks and Heppenstall, 2012), and make the outcomes of the models more understandable by the different audiences and decision makers (Wegener, 2021). However, the visualisation task is difficult due to the high computational demands of graphics (Heppenstall *et al.*, 2012). Also, experience and guided design are necessary to visualise the agent's behaviour in a communicative and cognitively effective manner (Kornhauser, Wilensky and Rand, 2009).

A.4.h Resources and software

Conventional programming languages were often used to develop ABM models in the early stages of research (Gilbert and Bankes, 2002). A significant disadvantage of this approach is that model developers would have to spend a significant amount of time and effort on procedures such as graphics, memory management, and synchronisation mechanisms (Chen, 2012). Authors with limited or no computer programming skills were largely discouraged by these implementation barriers to implementing ABM (Railsback, Lytinen and Jackson, 2006).

However, the use of simulation and modelling toolkits can greatly speed up the development of agent-based models. When it comes to programming parts that are not particular to a certain application such as a Graphical User Interface (GUI) and data import/export, toolkits allow modellers to focus on research and model testing rather than developing the basic tools required to run a computer simulation (Tobias and Hofmann, 2004). Also, given that professional developers have created and optimised complex parts as standardised simulation and modelling functions, this increases the model's reliability and effectiveness (Castle and Crooks, 2006).

Agent-based modelling tools are abounded. However, tools like NetLogo, AnyLogic, MATSim or Repast have been used frequently used in current studies and continually receiving software updates. Moreover, global Google searches for the software tools listed above shows that NetLogo is the most popular tool, followed by AnyLogic, MATSim, and Repast (Cenani, 2021). Castle and Crooks (2006) suggested that the following criteria be taken into consideration when selecting a toolkit: 1) ease of development of the model / use of the system; 2) size of the community using the system; 3) accessibility to help or support (most likely from other users) and; 4) size of the community familiar with the programming language in which the system is implemented. For model development in this study, NetLogo and AnyLogic were initially considered. NetLogo was considered given its considerable community size, ease of use, and the fact that it is free. Whereas, AnyLogic was considered because of the author's experience with the tool, its distinguishing feature of integrating the optimisation engine OptQuest and the developer's support. However, AnyLogic must be purchased and requires a subscription for support and maintenance.

To explain a little more about these models, NetLogo is a popular ABM platform for modelling complicated systems (Cenani, 2021; Railsback, Lytinen and Jackson, 2006).

Its popularity was gained from being an open-source multi-agent modelling platform with GIS support and movie recording (Cenani, 2021). Moreover, it is a self-contained tool that requires little programming expertise and includes modelling modules that let beginners do a lot with little code (Manson *et al.*, 2020). Due to the high number of users, sample models are abundant, adding to its appeal (Cenani, 2021). Railsback, Lytinen and Jackson (2006) argue that despite the fact that scientists may believe it to be too limited for use in serious ABMs, prototyping complex models with NetLogo is highly recommended. However, because NetLogo lacks an Integrated Development Environment (IDE) with features such as a text editor, resource monitoring, and etc., it might be challenging to create complex models (Manson *et al.*, 2020)

On the other hand, AnyLogic is a universal tool built to handle sophisticated models that could simultaneously integrate different types of models, namely Discrete Event (DE), ABM, and System Dynamics (SD) (Borshchev, 2013). AnyLogic offers numerous features beyond NetLogo, including database integration for data retrieval and writing, adjustable randomization functions, and libraries for generic modelling, such as the discrete-event library and subject specific such as the traffic modelling library. Importantly, it also includes an integrated optimisation engine. However, it is more expensive to maintain as proprietary software than open-source competitors such as NetLogo. Also, most of its source codes are not accessible biased the user write codes. The feature of its integrated optimisation engine, mainly led to AnyLogic being chosen as the simulation tool. At this point, it would be good to explain more about the OptQuest optimisation engine.

Before describing the optimisation engine in detail, it is crucial to mention that the subject matter of optimisation in-depth is beyond the scope of this research. Optimising the standards for education is a goal that is feasible via generic optimisation tools developed by experts in the optimisation field. This sub-section will touch upon the main ideas behind the OptQuest optimisation engine, and then the type of temporal problem associated with this research followed by a description of the inner working of the used software for optimisation.

For the main optimisation approaches related to OptQuest, it is important to understand the distinction between heuristic and metaheuristic optimisation. Meanwhile both approaches generate solutions that may not be the best possible but are sufficient given the constrained time window. The difference lies in the process of searching the solution space. Heuristic methods employ a set of rules that are tailored to the specific problem at hand (Pearl, 1984). In contrary fashion, metaheuristic approaches dynamically alter the search rules based on the results of previous optimisation trails. These rules are typically changed based on information derived from the optimisation objective and the variables used in the process in an iterative manner (Osman and Laporte, 1996). As a result of these optimisation attempts, metaheuristic algorithms have been adopted as a general-purpose search approach (Naghavipour *et al.*, 2022). According to Laguna (2011), the effectiveness of heuristic processes can be significantly boosted with the help of metaheuristics. This is due to the fact that the search strategies that are offered by metaheuristic techniques result in iterative processes that have the capability to explore solution spaces farther than the solution that is produced by applying typical heuristics. Moreover, during the cycle of its iterations, metaheuristics like Genetic Algorithms (GAs) and Scatter Search (SS) can optimise, based on a controlled set of solutions that they have previously generated, unlike Simulated Annealing (SA) and Tabu Search (TS), which often modifies a single solution in order to produce a new solution. See Osman and Laporte; Talbi (1996; 2009) for an in-depth discussion of metaheuristics.

Given that this research is targeting the next 50 years in optimising the provision standards for education, the optimisation is expected deal what is called Multi-periodic Optimisation problems. An optimization that works for one period (t) may not work for the next $(t + n)$ (Talbi, 2009). According to Fleten, Høyland and Wallace (2002), two common methods may be implemented based on a multi-optimisation time phases or a single time phase for the model time horizon, which might be referred to as dynamic and fixed, respectively. The optimising process is repeated for each stage in the dynamic method. This method may be more appropriate for period-responsive strategic planning with reduced time gaps. The fixed method, on the other hand, would give a single answer over the time horizon, possibly based on some type of averaging. In strategic planning interventions, this approach might be regarded as less responsive. However, fixed models can be more seriously utilised to address this type of problem (Talbi, 2009).

When it comes to "black box" optimizers, according to Laguna (2011), the OptQuest is a general-purpose "black box" optimizer that uses metaheuristic algorithms to generate an acceptable solution in a limited timeframe for complex systems. Furthermore, black box optimizers such as OptQuest are also known as context-independent procedures because they do not take advantage of the problem's specific structure, because many complex systems may lack a convenient mathematical representation, whether linear or nonlinear. This characteristic of independence enables it to avoid the necessity of formulating mathematical optimisation problems as linear, nonlinear, and integer. Figure 7 shows the relationship between the simulation model and the optimisation engine (i.e. Black Box).

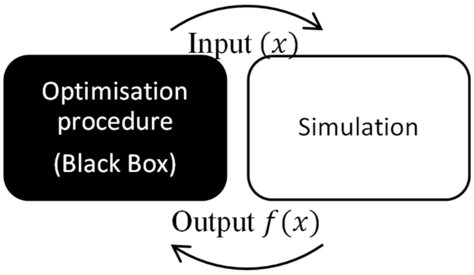


Figure 7 Simulation Optimisation Using a "Black Box" Approach (April *et al.*, 2003).

Figure 7 shows the optimisation engine is in charge of providing the simulation model with the inputs for its variables, which are denoted by x. After that, the output of the simulation, which is denoted by $f(x)$, is provided back to the optimisation engine so that it can be evaluated and a new set of inputs x may be generated to be tested. To provide further clarification about the inputs and outputs, Laguna and Martí (2002) state that OptQuest was developed to search for optimal solutions for a certain class of optimisation problems, which includes the following:

Max/Min:

$$f(x) \qquad \text{(Simulation output)} \tag{9}$$

Subject to:

$$\begin{aligned} &l \le x \le u && \text{(Bounds)} \\ &Ax \le b && \text{(Constraints)} \\ &g_l \le g(x) \le g_u && \text{(Requirements)} \end{aligned}$$

where $f(x)$ is the simulation function of the variables x that must be assessed for maximisation or minimisation. Therefore, for each variable x, a lower and upper bond value is assigned, denoted by l and u correspondingly. A set of constraints Ax for the variables x may be specified linearly with respect to given values of b. Just before simulation, the optimisation engine examines these Ax restrictions to filter the newly generated set of inputs for operating $f(x)$. In addition, some requirements may be applied, as shown by $g(x)$ as a result of the subfunctions for the variable x. After the simulation is complete, these required values are verified against the values of the lower bound g_l and the upper bond g_u, both of which must also be known in advance. Even though this approach is described to be context-independent, Laguna (2011) explains that the context-independent concept is more difficult to define since no solution is completely independent of the context. This is because information about the context may be partitioned between the objective function f(x) and the defined restrictions which are feedback into the optimiser.

To provide an overview of the OptQuest workflow, it can be stated that it consists of three primary components: an SS Algorithm for space exploration, a TS algorithm for solution filtering and suggestion, and an NN Accelerator. Both Glover, Kelly and Laguna (1996) and April *et al.* (2003) explain the technique as follows: initially, the solution space is investigated using the SS approach. This algorithm is a member of the class of population-based algorithms that generate multiple solutions to a problem. SS comprises of the five steps listed below: 1) the diversity generating strategy is used to generate a group of diverse solutions that serve as search starting points; 2) the improvement strategy improves solutions with the aim of boosting quality, which is typically measured by the value of the objective function; 3) the method for updating the reference set constructs and maintaining a collection of solutions for the primary iterative loop of scatter search; 4) the subset generating method which produces new subsets of reference solutions as inputs to the combination method; and 5) the solution combination technique which generates additional trial solutions based on the subset generation method's result. In general, new trial solutions result from the combination of two or more reference solutions. The second approach is TS, which generates a single solution to filter the SS results. The TS algorithm, unlike the SS approach, is a single-based solution. In addition, TS employs a memory element to prevent iteration of the same solution and increase the attractiveness of unexplored regions of the solution space. The third phase involves the addition of a NN that is trained throughout the optimisation procedure to "filter out" particular variables of x that are likely to result in an excessively low value of $f(x)$. This comprehensive description of the workflow may not cover all of the possible OptQuest paths in an optimisation task due to the possibility of switching to a linear or mixed-integer programming solver based on the results of other integrated evaluations in OptQuest (Laguna, 2011). Nevertheless, it does cover the basic optimisation pathway in OptQuest.

B. A comprehensive example of contemporary models

O'Donoghue *et al.* (2013) argues that UrbanSim is one of the most powerful and well-known modelling frameworks available. Moreover, it has been used in several places across the world to anticipate urban features. Among the significant characteristics that persist in UrbanSim are: 1) High modularity: UrbanSim consists of multiple sub-models that might be separately developed or exchanged without affecting the other sub-models. This is achieved by using a shared data store and a dedicated coordinating mechanism (Waddell *et al.*, 2003); 2) Highly disaggregated: considering parcel level land-use, more than 100 types of households, explicit business establishments, and explicit policies integration (Hunt, Kriger and Miller, 2005); 3) Dynamic disequilibrium: each sub-model can have its own updating times executed independently from the other sub-models (Waddell *et al.*, 2003); 4) Clear behavioural realism: UrbanSim oriented opposed to the use of abstract modelling assumptions that are not reflected in observed behaviour (Feldman *et al.*, 2010); 5) Large scale and long-term: while utilising a high level of disaggregation, UrbanSim simulates the potential long-term effects of various policies on major urban regions (Schwartzman and Borning, 2007); and 6) runs on standard computers: UrbanSim can be used on modern laptops, with different operating systems and with reasonable speeds. For example, using a full population for microsimulation, a parcel-level model of San Francisco can take about 20–30 minutes for every 10 simulation years (Feldman *et al.*, 2010).

However, UrbanSim can be criticised for being: 1) Data hungry: as calibrating a model to be fully operational requires a significant amount of data to be collected, and processed (Nguyen-Luong, 2008; Zhao and Chung, 2006). This aspect limits the use of individual choice models (logit models) and typically replaces them with trend-based models; 2) Over-perceived generalizability: applying the model outside of the United States urban context necessitates significant alterations (Nguyen-Luong, 2008); 3) Requires an inter-disciplinary group, as well as the assignment of sufficient resources. (Nguyen-Luong, 2008); 4) UrbanSim is not an explicit ABM but, it has some features that can be described as agent-based (Hunt, Kriger and Miller, 2005). For example, there is no interacting representation of the population individuals or their households as mobile agents; 5) The model outcomes are mainly Indicator based visualisations: with the help of indicators, the simulation results are presented in a clear and concise manner (Schwartzman and Borning, 2007). Moreover, the urban environment such as land parcels, building blocks and streets layout are presented similar to typical three-dimensional GIS views. However, it lacks the transparency of processes gained from animated agents such as those in transportation or population movement in the space; and 6) Nguyen-Luong (2008) believe that UrbanSim development was dominated by computer and economics experts rather than urban planners and modellers. Despite the generic problems associated with operating a complex systems that applies to UrbanSim and any similar framework such as being data hungry, UrbanSim is continuously being improved on the technical and practical side (Nguyen-Luong, 2008).

In order to understand how UrbanSim operates, both the software architecture and urban system components of UrbanSim will be explained in this section. According to Waddell *et al.* (2003) the software architecture consists of four major components: 1) the sub-models comprising the urban system; 2) a model coordinator that schedules the execution of sub-models; 3) an Object store (i.e. information storage) accessible by the various models for information exchange; and 4) a mediation layer to aggregation and transform the data before being exchanged among the sub-models. See Figure 8 for the software architecture of UrbanSim.

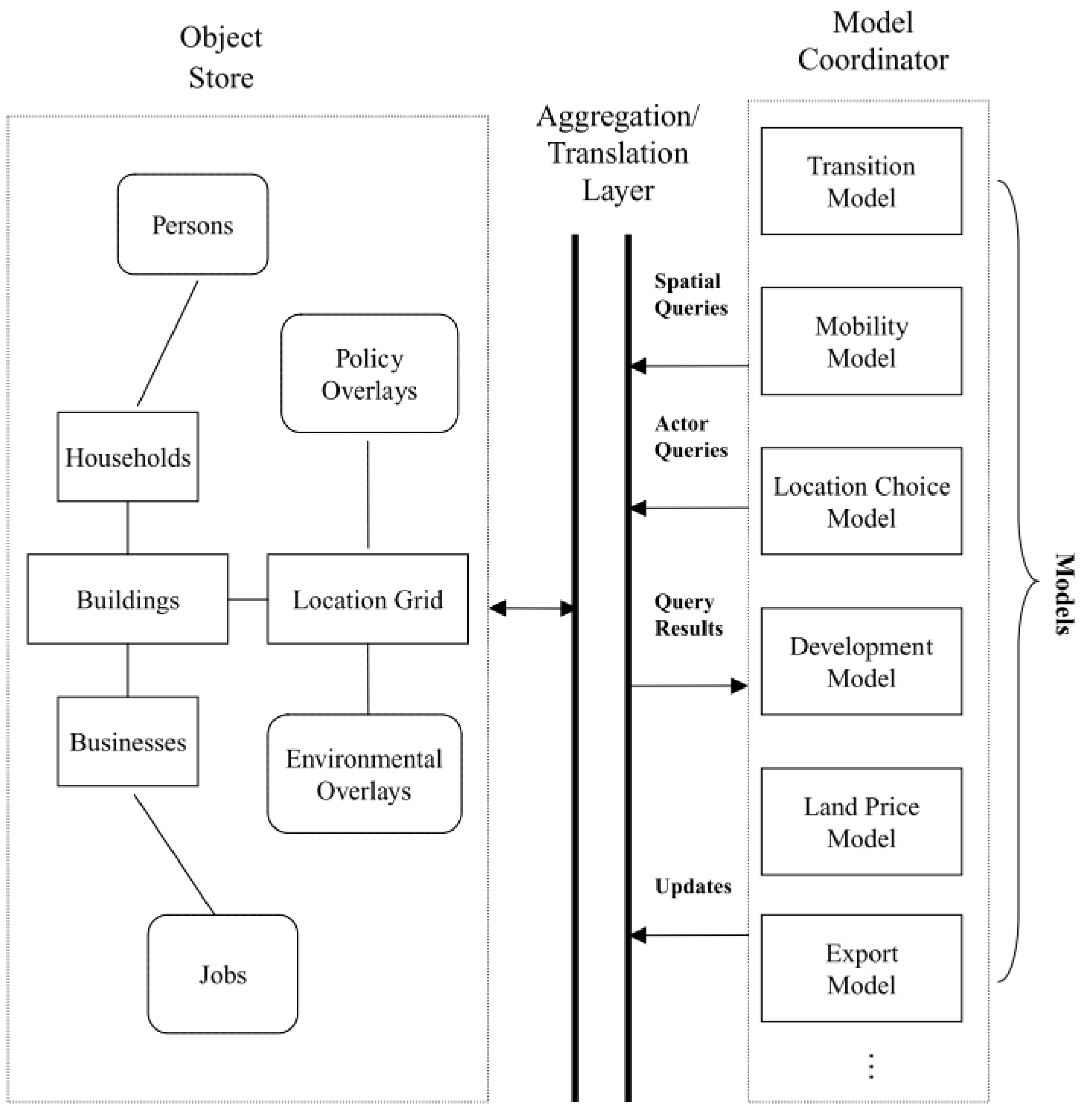


Figure 8 The software architecture of UrbanSim adopted from (Waddell *et al.*, 2003).

Figure 8 shows the sub-models of the urban system, which are situated under the Model Coordinator component on the right-hand side of the diagram. On the other hand, on the left-hand side is the object store where the various data collections about the population, business and environment are stored. In the meantime, the data updates, queries, and results for the sub-models are managed by the Aggregation and Translation layer, which are located in the middle of the figure.

Regarding the urban system components in UrbanSim it consists of the following (Waddell *et al.*, 2003; Zhao and Chung, 2006):

1) The Accessibility Model creates indices of accessibility from each district to the different activities by travel mode that are considered vital in the decisions of locating household and business. A logsum function is utilised in this model to calculate the accessibility weights.
2) The Demographic Transition Model simulates the population change in terms of, aging, household formation, births, deaths, migration, income, and etc. This sub-model uses microsimulation modelling and Iterative Proportional Fitting (IPF).
3) The Economic Transition Model simulates employment change in terms of jobs increase and decrease by sector. This model also utilises microsimulation modelling and Iterative Proportional Fitting (IPF).
4) The Household Mobility Model predicts whether a household will relocate based on historical trends. The decision to relocate includes the new residence type and the district name.
5) The Employment Mobility Model predicts which jobs will be relocated throughout a given year based on historical trends also.
6) The Household Location Model determines which place each household moves to based on the available residential units in the targeted district. The selection of the final location process incorporates sampling the vacant residential units and a Multinomial Logit decision function.
7) The Employment Location Model determines where to place relocating and newly established jobs. The modelling approach is also based on sampling and the Multinomial decision function.
8) The Real Estate Development Model replicates the decisions developers make about the location and the kind of projects to build. The model runs over all built and non-built grid cells (i.e. the metropolitan land) where development is permitted and provides a list of possible project alternatives, including the option of not developing different areas. A multinomial logit model is used to estimate the likelihood of each alternative being selected. The model inputs include the

attributes of the site and its surroundings such as the current building status if the land is built and its proximity to existing developments.

9) The Land Price Model simulates the change in land prices by utilising a hedonic regression approach that considers the land and its surroundings. The considered aspects may include aspects such as the accessibility levels to the district and the effect of imposed building regulations on land prices.

According to Waddell *et al.* (2003) In addition to these sub-models, UrbanSim could take inputs from external models such as:

1) The Macroeconomic Model, which forecasts future macroeconomic circumstances like population and employment by sector; and
2) The Travel Demand Model, which estimates travel conditions such as congestion of travel across zones.

See Figure 9 for an overview of UrbanSim sub-models and sequence of execution adopted from (*Waddell et* al., 2003).

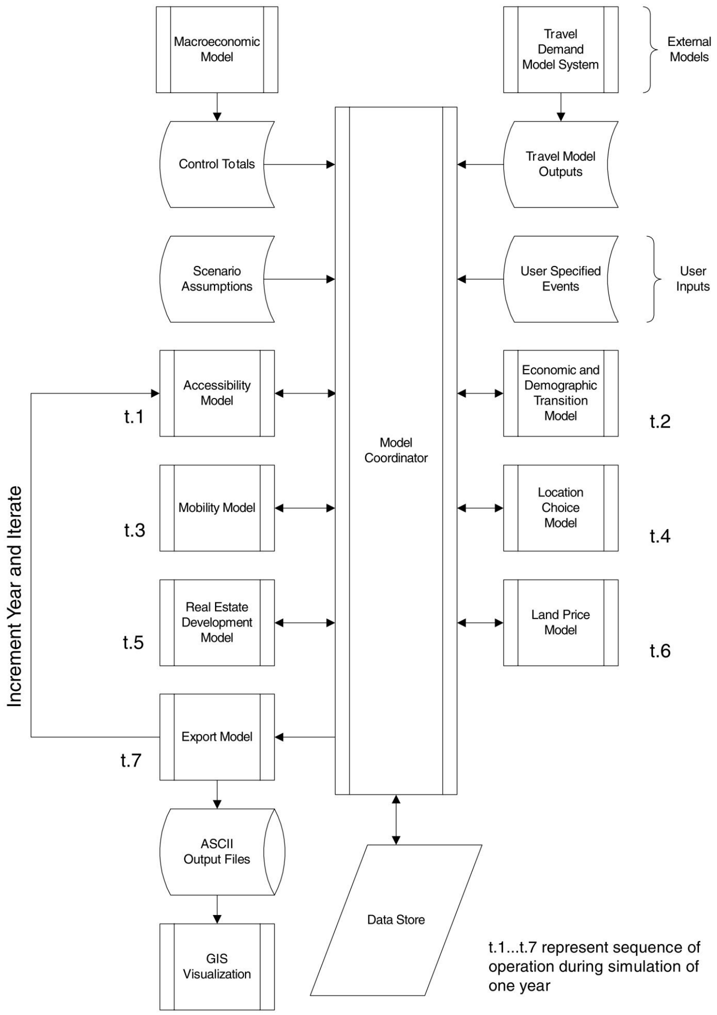


Figure 9 An overview of UrbanSim sub-models and sequence of execution adopted from (Waddell *et al.*, 2003).

Figure 9 shows the UrbanSim main framework that includes the urban system sub-models within the general software architecture. The main sequence of executing the sub-modes is numbered from (t.1) to (t.6) for every iteration. Moreover, the external inputs include the User inputs in the form of selecting the modelling events and the projection assumptions. Additionally, the Export Model that writes the simulation outputs to user-specific format (ASCII output files) for additional analysis or processing such as GIS visualisation. Therefore, this component is not a model in the traditional sense as it exports the model outcomes.

UrbanSim can be viewed as a successor to the Lowry model described earlier in the traditional Planning Support systems given its modularity. Moreover, UrbanSim and the similar comprehensive models such as IRPUD (Wegener, 2011) and HUDS (Kain and Apgar, 1985) are significant milestones for the progress of Planning Support Systems. However, the limitations imposed by the complexity of the urban system, the lack of knowledge in the related subject matters, the scarcity of data availability, the lack of documentation in terms of the tools and methods applied and the practical examples to learn from and the gap in the urban planner's participation in the development of planning aid tools, all contribute to its limited use. This research will not only showcase the process of building a similar tool from scratch but aim to engage more planners with its direct implementation to the issue of calibrating the land use provision standards.

2.3 Urban planning aid tools In Saudi Arabia

Various government agencies in Saudi Arabia have adopted or attempted to implement appropriate models for specific applications of their interest. The author is familiar with two models aimed at providing educational facilities, either directly or indirectly, as well as a third model that is more general in scope but is used for energy studies instead.

2.3.1 The ADA location-allocation model for public facilities.

As part of the action plan for the provision of public facilities in Riyadh, ADA created a location-allocation model to improve the standards for providing public facilities. The model was developed incorporating joint work between ADA and the private consultant Khatib & Alami. Moreover, the model is implemented using ArcGIS model builder.

The ADA action plan for public facilities is comprised of several steps that can be categorised into three major phases: first, upgrading the city database to include newly added land parcels and the built environment attributes; second, an evaluation of the operational level of public facilities. This stage is performed with the assistance of an automated tool that assesses the degree of conformity between standards and their actual application; third, the revision of standards and the development of a time-bound action plan. Due to the limited importance of the first step to the allocation model's workflow, only the second and third stages will be described in detail (HCDA, 2017).

The evaluation of the standards conformance is done according to the following steps: 1) allocating the population at the parcel level; 2) Generating catchment area polygons for each facility location; 3) Using the created catchment area for each facility to get the average walking distance and covered population; 4) Using the suggested standards as benchmark values for service evaluation (i.e. the evaluation of conformance); and 5) the visualisation of results using heat maps. See Figure 10 for an example of the ADA standards conformality evaluation outcome adapted from AlSaif (2016).

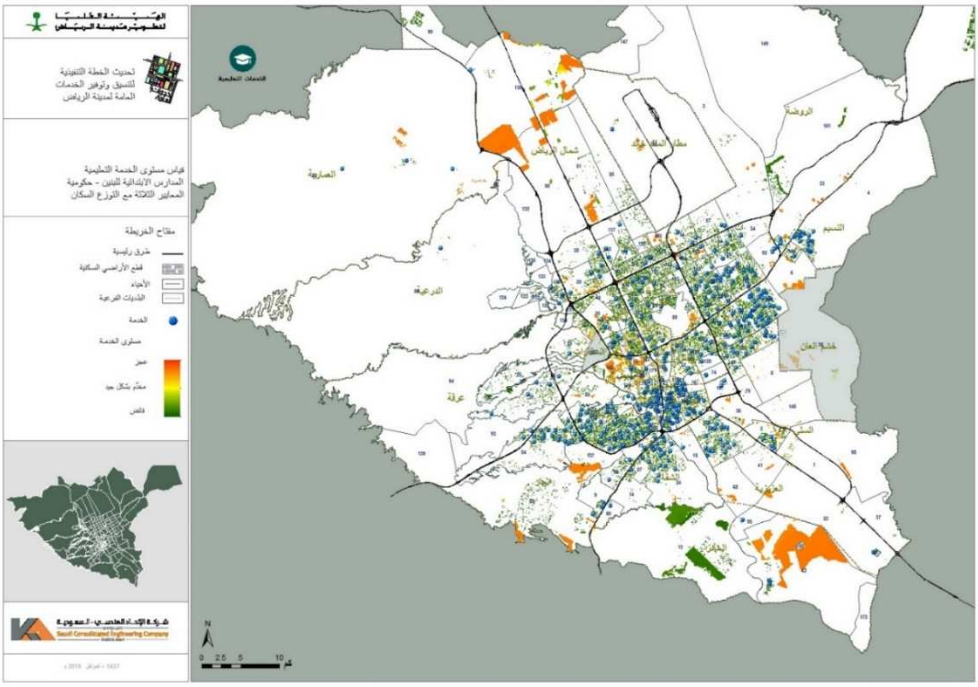


Figure 10 an example of the ADA standards conformality evaluation outcome adapted from (AlSaif, 2016).

Figure 10 shows a heatmap representing overlays of population coverage, average walking distance, and the defined area per person for elementary education in Riyadh. This model considers the demand is proportional to population given the predefined standards’ criteria.

The complete workflow of the location-allocation model including the specifics of population growth estimation, or the population relocation modelling are not known to the author. However, the model can be applied to the current or future year based on given population estimates at the parcel level. Moreover, the calibration workflow and its objective for enhancing the standards can be described as follows:

Given the standards evaluation outcomes from the standards conformality tool, there are two pathways. First, if the supply levels are assessed to be higher than the standards then, the standards should be lowered only if found to be higher than the global average. On the other hand, if the supply levels are evaluated to be lower than what the standards defines them to be, the standards will be compared to the global practice and either kept 'as is' or lowered if possible. After this process the standards are reformulated to enhance the supply levels. It can be noted from this process that improving the standards does not include rising the standards (ADA, 2011).

The outcomes of the allocation model provide further details about the suggested location of the specific facilities to be built in year 2020 at the district level and the numbers of additional facilities every five years up to the targeted year. Overall, the solutions suggested by the ADA action plan for tackling the public functions problems involved other tactics based on enhancing the standards such as: 1) land acquisition when possible, estimated to cost 13 billion Saudi Riyal, or collaboration with private parties in different models; 2) enhancing the available database; 3) preventing any further land use transformation to other uses; 4) phasing the action plan into four phases to the year 2030; and 5) prioritizing the function provision (ADA, 2011).

2.3.2 The Other Ongoing Models

During the search for data for this research, the author became aware that around 2017 the King Abdullah bin Abdulaziz Public Education Development Project started working on a location-allocation for provision of schools. However, the case study for that project was Makkah City. Unfortunately, no further information is available about the project progress or outcomes.

On the other hand, in 2020 the King Abdullah Petroleum Studies and Research Centre (KAPSARC) started working on an activity allocation model for Riyadh using the Production, Exchange and Consumption Allocation System (PECAS). The activities included in the model include labour market, real estate market, and trips on the transport network. However, the model is utilised in the context of spatial urban energy systems instead of urban planning. According to Hunt (2003), the PECAS model is mainly composed of two modules. First, the developer's model which provides space area via new development, demolition and re-development for activities to locate. Second, the activity allocation module, which locate the specific activities targeted in the area supplied by the developers' model and how these activities interact with one other at a particular time. Moreover, households are included in different types and categories based on economic activities. When it comes to labour, it is a commodity that is generated by households and consumed by the economic activities. Fundamentally, the model employs a short-run equilibrium based on development event occurrences that is generated using some kinds of Logit model (Hunt and Abraham, 2009). Although this model does not serve the purpose of this research explicitly, it belongs to the family of urban models that can be calibrated to serve the purpose of providing both commercial and public facilities as economic activities.

It is also crucial to note that additional planning support systems or tools are utilised in Riyadh, such as transportation models that try to address issues such as traffic congestion and the estimation of daily journeys, and which may be incorporated as sub-models into more complete urban models.

2.3.3 Remarks

Despite having general education being optimised in the case study to be developed here, the literature review has covered both general education and commercial activities. While the results of this study could be used to improve the existing standards for public services. The allocation standards for commercial activities are still lacking. Moreover, the privatisation of public schools can demonstrate the planning support system's applicability to both public and commercial providers. This is because the similarities between the two providers lies in: 1) the way the educational system is organised around a specific population demography for different stages that are relevant for consumer profiling in any commercial activity; 2) the utilisation of similar ratios per user or typical building designs in the provision of services which is directly feed into the provision cost benefit analysis; and 3) the competition aspects among the different land uses for land utilization that are found from general education which normally consists of elementary, middle and secondary stages distinguishing between separate or mixed gender schools. As such, the system could demonstrate the simultaneous optimisation of multiple land uses within a single land use category.

Chapter 3

Data

This chapter will address the most important aspects of data in terms of its significance, quality dimensions, sources in Saudi Arabia, main classification and availability, data requirements, and timeliness. Then, based on the classification of the data as pertaining to population data and built-environment data, the data will be rigorously described, analysed, and processed to highlight the challenges and actions necessary to prepare the data for the modelling stage of this project.

3.1 Data importance

The data aspect in urban modelling has great relative influence on the outputs of the various models. This is because input data directly impacts the models' outputs independently of which methodology is being used. A methodological change in urban simulation may aim to enhance aspects such as the model accuracy in calibration, speed of execution, level of realistic imitation and sensitivity to variation but it is not intended to fix core information or trends that are carried by data. Therefore, the usable data must be of high quality, despite the fact that the definition of data quality (DQ) can be as simple as "fitness for use" (Tayi and Ballou, 1998). The assessment and handling of this matter is multifaceted in consideration (Wang and Strong, 1996), user-centric in handling (Tayi and Ballou, 1998) and lacking in terms of meaningful knowledge and structured processes (Clarke, 2016)

3.2 Data quality dimensions

According to Wang and Strong (1996), DQ can be forked into four main categories as Intrinsic DQ, Contextual DQ, Representational DQ and Accessible DQ with a large number of dimensions to consider. To explain more about these angles to DQ, then we need to note:

Intrinsic DQ focuses on aspects such data accuracy and credibility.

Contextual DQ considers the fitness of the data for the task at hand like the added value of the data, timeliness, and amount of data observations.

Representational DQ deals with aspects such as legibility of data, data formatting and ease of manipulation, and

Accessibility DQ is concerned with the rights of usage and the mediums of storage and retrieval.

3.3 What affects DQ and how it reflects on the process of collecting the representing data

One of the main issues affecting data quality is the separation between primary collectors and consequent consumers. This split causes the data to be repurposed by its new users, limiting the control over the data quality which is normally evaluated during the collection process and diminishes the access to its metadata (Stein, 2007; Tayi and Ballou, 1998; Zhang, Indulska and Sadiq, 2019) Consequentially, data collection as a process would be inverted from its well-known top-down design guidelines into an ad-hoc task that does not follow an established process aiming to explore and resolve the data quality in a bottom-up manner (Zhang, Indulska and Sadiq, 2019). Therefore, there is a substantial responsibility on urban modellers who use repurposed data to ensure the data suitability for their intended use. Moreover, it is not surprising to find that the data collection and preparation task for urban simulation models is known to be as difficult and time-consuming task (Lee, 1973; Timmermans, 2003; Waddell, 2011).

3.4 The main data sources in Saudi Arabia

In Saudi Arabia, the General Authority for Statistics (GAS) is the main body used to collect and reference official data and national statistics. However, the reach across other government entities is essential for covering any lack of data. Moreover, due to the coverage size of this research which considers all districts of the metropolitan city and the required level of disaggregation in information, data gathered by Individuals and small research groups has become scarce and hardly provides any added value to be considered. For example, information about each school capacity in Riyadh city was only available through the MOE. Moreover, the only source for residential mobility data for this research was in the ADA demographic survey. In total, seven government bodies were identified and contacted during the data collection process, and these were as follows:

1) GAS, is the official body to collect, analyse and publish data In Saudi Arabia. Census and demographic surveying data are the main themes of data used from this source.

2) Riyadh Municipality (RMU) is the main administrative and supervising body for managing Riyadh city. The acquired data from this source is related to building and commercial permits.

3) ADA, the planning arm of Riyadh Municipality. This entity conducts its own planning related surveys and studies to support their planning decisions. The data acquired from ADA includes Riyadh's GIS base maps and demographic survey data for the year 2016.

4) Ministry of Health (MOH), beside its role of providing health care, is a main recorder of births and deaths in Saudi and the data is published in the MOH annual reports.

5) MOE, given its main role in to education, has a large data base of all educational facilities and enrolled students. The provided data from MOE consisted of a list of all schools in the Riyadh region with their attributes such as number of classes, capacity for students, and building condition. Moreover, a list of all available educational allocated land in Riyadh city and additionally, the numbers of all enrolled students in all education stages for the last two decades are available.

6) Ministry of Justice (MOJ) is the authority responsible for documenting property exchange. An access to properties' transactions records was available which enabled access to a detailed information about properties' exchange such as price, area and district location.

7) King Abdulaziz City for Science and Technology (KACST), the scientific government entity to support and enhance applied research, has several data bases. From its wide range of research themes, limited access to data about Riyadh daily travel behaviours and the Kingdom Internal migration trends is possible though their published work.

The publication of data by Saudi government entities rarely adheres to the FAIR data standards (Wilkinson *et al.*, 2016) as promoted by UCL because the data and its metadata cannot be found or accessed using unique identifiers. Yet, the data is moderately accessible and interoperable.

3.5 Data classification and geographical availability

Due to developing two separate models to deal with the city and its districts, the amount of required data that is disaggregated demographically and geographically became an obstacle. This is caused by various reasons such as lack of collection at small area level or not being published due to strict data policies. To overcome this common issue in urban modelling, different techniques could be used such as data propagation, synthesis and aggregation. For example, due to the lack of city level data, the fertility rates used at the city level model belongs to the Kingdom level. Therefore, this sub-section is meant to clarify the data sources and relations under its use classification and geographical representation. Moreover, this helps in identifying the data shortcomings and the required data to improve the models' inputs. Figure 11 shows the data classification chart by intended use at different geographical levels.

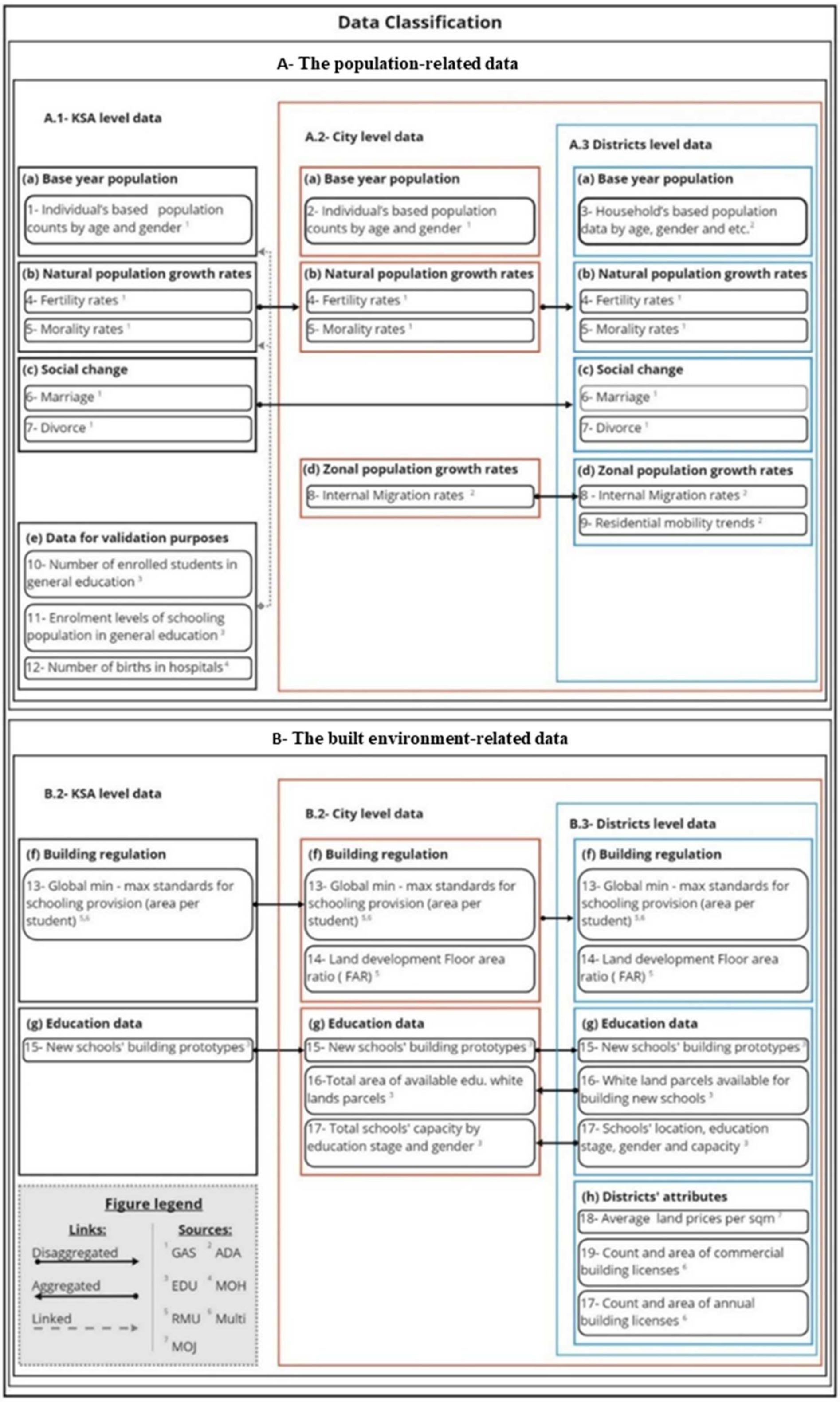


Figure 11 Data classification chart by intended use and geographical levels.

Figure 11 shows that that data can be mainly classified into population-related data and built environment-related data. Each of these two groups is further divided into three geographical levels as the Kingdom level, the City level and the Districts level. This classification helps in understanding the geographical extent of data availability and whether the data has been used directly in its geographic boundaries or imputed from a different geographical level. The population related data is grouped into five main categories: 1) Base year population (census and demographic surveys); 2) Natural population growth rates (fertility and mortality rates); 3) Social change rates (marriage and divorce rates); 4) Zonal growth rates (migration and residential mobility rates); and 5) Validation data (such as the number of enrolled students, enrolments rates in education stages and birth numbers in hospitals). The built environment-related data were grouped into three main categories as: 1) Building regulation (floor area ratio and schools provision limits (min max)); 2) Education data (Building prototypes, available white lands and attributes of built schools); and 3) District attributes (average land prices and residential and commercial licenses attributes).

By noting the directions of interconnection between geographical levels in Figure 11, approximately one third of the data used were adopted from the upper level of the Kingdom to smaller levels. For example, due to lack of natural growth rates and social changes at the city and districts level, it has been propagated from the Kingdom level. Also, it should also be noted that there is a vertical link for the data used for validation purposes at the Kingdom level. This is because there are some disturbing observations in both population and birth rates, which have forced us to look for additional sources to verify the quality of the data and adjust it as possible. The observed problems and ways to address them will be addressed in the following sub-sections.

3.6 The required data

The required data for the developed models at the city and the districts level are different in terms of their inputs and disaggregation. For example, while the city level model uses total white land areas available for educational purposes, the districts level requires a list of all white land parcels available for educational purposes by location and area. Moreover, both city and district models contain at least two sub-models with different inputs. Figure 12 illustrate how the data is used within these different models.

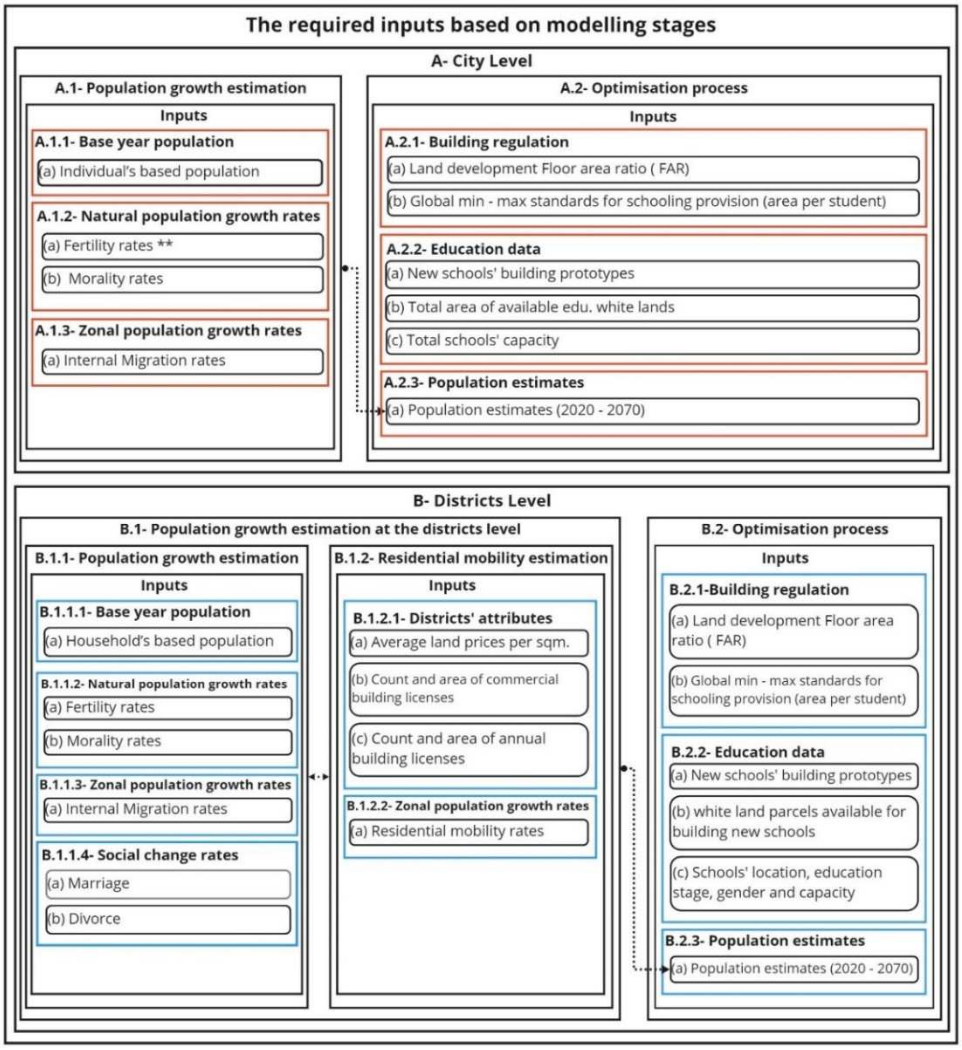


Figure 12 The required data based on modelling stages.

Figure 12 shows that the data required at the city level can be divided into two groups according to the model process. These processes are (A.1) the Population growth estimation and (A.2) the Optimisation process. The data associated with population growth estimation can be grouped into 5 categories as Base year population, Natural population growth rates, Zonal population growth rates, Social change rates, and Districts' attributes. The data used in the optimisation process are classified into 3 groups as Building regulations, Education data and Population estimates. The input data of population estimates in the optimisation process represent the outcomes from the population growth estimation process as shown by the linking arrows. It is clear from Figure 12 that 18 different types of data are needed to run two study models, 10 of which are for the city level model and 16 for the districts level model. The fundamental difference in the needed data for the two models lies in the additional requirements for the population forecasting phase in the districts model compared to the city model. In general, all data are required with extensive amounts of detail and disaggregation. In addition, rates of natural growth, inter-regional migration, social changes, and districts attributes are naturally dynamic variables. Therefore, the possibility of anticipating future changes for these rates should be determined based on the goodness of available data as well as the theories/knowledge available for each factor.

3.7 The timeliness aspect of available data

The time factor is considered as one of the most important factors affecting the data. This is because it allows us to know the proximity of the available data to the events to be studied and because of the restrictive time considerations for the possibility of using the data in future projections. To clarify what is meant by restrictive time considerations, most data forecasting models require the availability of a considerable number of historical data points, with equal time intervals in between data, such as every five years (2000, 2005, 2010, 2015, ... and so on). This is because it is more difficult to deal with varying time intervals between reference data which would limit the possibility of forecasting. Given the fact that various datasets were used, and collected from different sources, the timeliness factor had to be considered in an approximate manner. To clarify the process of arranging the data chronologically Table 4 shows the time groups classification, year of data, the description of the data and its sources.

Table 4 the time groups classification for available data.

Years' Group	Year of data	Data description	Data source
(Pre year 2000)	1975 - 1995	Fertility rates[1]	GAS (2004,2007,2016)
	1991 - 1995	Residential mobility[1]	ADA 2016
(Base year 2000)	1995 - 2000	Fertility[1]	GAS (2004,2007,2016)
	1996-2000	Residential mobility[1]	ADA (2016)
	1999	Demographic survey (F, M)	GAS (1999)
	1999-2001	Count of enrolled students[2]	MOE (2018)
	2000	Count of births in hospitals[2]	MOH (2018)
	2001	Demographic survey (F, M)	GAS (2000)
(Year 2005)	2000 - 2005	Fertility rates[1]	GAS (2004,2007,2016)
	2001-2005	Residential mobility[1]	ADA (2016)
	2002 - 2006	Count of enrolled students[2]	MOE (2018)
	2004	Census (F, M)	GAS (2004)
	2005	Count of births in hospitals[2]	MOH (2018)
(Year 2010)	2005 - 2010	Fertility rates[1]	GAS (2004,2007,2016)
	2006-2010	Residential mobility[1]	ADA (2016)
		Count of births in hospitals[2]	MOH (2018)
		Count of enrolled students[2]	MOE (2018)
	2007	Demographic survey (F, M)	GAS (2007)
	2010	Census	GAS (2010)
		Average land prices per sqm[2]	MOJ (2018)
		Licenses (C, R)[2]	RMU (2018)

Table 4 the time groups classification for available data (continued).

Years' Group	Year of data	Data description	Data source
(Base year 2015)	2010 - 2015	Fertility rates[1]	GAS (2004,2007,2016)
	2011 -2016	Residential mobility[1]	ADA (2016)
		Count of enrolled students[2]	MOE (2018)
		Count of births in hospitals[2]	MOH (2018)
		Licenses (C, R)[2]	RMU (2018)
		Average land prices per sqm[1]	MOJ (2018)
	2016	Demographic survey (M, F)	GAS (2016)
		Demographic survey	ADA (2016)
Iteration years (2020 - 2070)	2017	Count of births in hospitals[2]	MOH (2018)
	2017-2018	Count of enrolled students[2]	MOE (2018)
		Licenses (C, R)[1]	RMU (2018)
		Average land prices per sqm[2]	MOJ (2018)
		Available white lands for residential development	ADA (2018)
	2018	Schools' building conditions[2]	MOE (2018)
		Education white lands[2]	MOE (2018)
	2019	Marriage and divorce[2]	GAS (2019)
	2020 -2070	Fertility rates[3]	To be estimated
		Licenses (C, R)[3]	
		Average land prices per sqm[3]	
		Population counts[4]	

[1] data originated in recent data collection
[2] data originated in the same year of data representation
[3] data forecasted in advance using sub-models.
[4] data forecasted by main models
(F, M) fertility and Mortality data
(C, R) commercial and residential data

Table 4 shows the division of time groups into 6 groups with five-year time periods, starting from the years pre-2000 and ending at the group iterations of years 2020 to 2070. To explain more about these time groups, each group was named after the primary year to be represented and the years included in each group were not strictly specified. The previous five years, including the year in which the group's label is named after, are mainly the constituents of each group. For example, the year 2000 collection covers the period from 1995 to 2000, but data from very recent years can be added if it is necessary for the time group. For example, the 2016 demographic survey was included in the group of year 2015 instead of 2020 for its necessity as a base year population data. The time groups can be categorized into three types, namely base year, inter-years, and future iteration groups.

For the base year groups identified for the different models, these were chosen based on the completion of the data and the time dimension that enables us to review the outputs of the models. Moreover, all base year data utilised here exclusively includes Saudi citizens of all ages and is consistent across all datasets. These years are 2000 and 2015.

In general, the year 2000 was chosen for the sub-models as a basic year from which the models start at the earliest point available for population data. For the year 2015, it is the base year from which the simulation of the main approaches start because the data is also available and is the last available time point of population data. For intergroups, they are of equal periods around the base years and contain the same amount of data to be used as inputs and verifying points for the reliability of the model's outputs as a pre-2000, 2005 and 2010 group. The last category of future iteration years are the groups between 2020 and 2070. This group represents the beginning of projecting all data required in all modelling levels. It also contains data that are subject to change depending on the outputs of other models, such as the amount of land available for the construction of residential and educational buildings.

It is worth noting that the year represented by the data and the history of the source show that there could be four different cases of data availability:

> 1) The existence of rates that have been extracted for previous years through the collection of more recent data. In other words, these data were not collected in the same year but were extracted from recent data to reflect earlier periods of time, such as residential mobility data and fertility rates data. In the case of residential mobility, the data collected in 2016 show the length of living in the last residence and the year of moving out from it, through which the trends of the movement of the population in previous years were known by aggregating the data by the year of the move in every past five year periods.
>
> 2) The existence of data collected in the same year it represents through records or surveys such as census data, real estate exchanges, enrolled students and white land available to the education sector, is available
>
> 3) Futuristic data that is projected independently using sub-models. These data are used as inputs for both the city model and the districts model. An example of these data could be future birth rates, average land prices and building permits.
>
> 4) Futuristic data that are a product of the core models to find the best solution for education provision standards.

Some data contains more than one type of information, such as demographic surveys and censuses data that would contain birth rates, death rates and migration data, as well as license data that contains information about business and residential activity. This data is embedded and encoded in tables for the purpose of simplification. In this way, the data were grouped and considered at fixed time points to make up for the lack of data from the different entities and to enable the operation of such demanding models as close as possible to equal reference time gaps.

3.8 The data in details

3.8.1 Population related data

3.8.1.1 Introduction: Basic population data exploration

Population related data are available in two forms Census and Demographic surveys. Since the establishment of the GAS, it has carried out 5 Census and 4 Demographic surveys. Also, the ADA established 5 demographic surveys. Therefore, theoretically, it can be said that there are about 14 comprehensive population related data sets available for Riyadh city. However, the only source that targets the households as a unit of population at the districts level in Riyadh city is the ADA. For the task of collecting population data only in 7 data sets, these were partially accessible as follows:

- From GAS:
    - a. Census data
        - i. 2004
        - ii. 2010
    - b. Demographic surveys data
        - i. 1999
        - ii. 2001
        - iii. 2007
        - iv. 2016
- From ADA:
    - a. Demographic surveys data
        - i. 2016

To assess the quality of these comprehensive data and their use of sub-data in the research, each subset will be detailed separately in the following sections.

3.8.1.2 Population based on individuals data (city level)

A. Description of original data format at the city level

Population data at the level of individuals is used in the city level model. This data is simple in terms of the amount of detail as it represents the total count of certain demographic characteristics such age and gender as shown in Table 5.

Table 5 the demographic classification of population data at the city level.

Age group		Gender	Total counts per age group
0-4	5-9	Male Female	Counts
10-14	15-19		
20-24	25-29		
30-34	35-39		
40-44	45-49		
50-54	55-59		
60-64	65-69		
70-74	75-79		
80+			

Fertility rates, on the other hand, are available in two forms as Age Specific Total Fertility rates (AS-TFR) and Age Specific Complete Fertility rates (AS-CFR). The AS-TFR covers women in childbearing years and provides information about the rates of annual births for each age group within the childbearing years. The data format is illustrated in Table 6.

Table 6 AS-TFR original data format.

Mother's Age-group		Gender of live birth	Number of births by mothers' age group	AS-TFR
15-19	20-24	Male Female	Counts	Ratio
25-29	30-34			
35-39	40-44			
45-49	50-54			

The AS-CFR covers women who have ended their childbearing years. This ratio represents the average number of children per woman during her lifetime. While, the ratio of women aged between 15 and 55 would be considered incomplete, the elder age-groups are considered as complete. The available data format is shown in Table 7.

Table 7 CFR original data format.

Ever-married women Age group		Gender of live birth	Number of All live births		AS-CFR
15-19	20-24	Male Female	0	1	Ratio
25-29	30-34		2	3	
35-39	40-44		4	5	
45-49	50-54		6	7	
55-59	60-64		8	9	
65 +					

Moreover, mortality rates covers all age-groups and for both genders. The available data is available as follows in Table 8.

Table 8 Mortality rates original data format.

Age group		Gender	Survival rate per age group and gender
0-4	5-9	Male Female	ratio
10-14	15-19		
20-24	25-29		
30-34	35-39		
40-44	45-49		
50-54	55-59		
60-64	65-69		
70-74	75-79		
80+			

For the city level model, the internal migration data has been used. However, it should be noted that, by and large, data about zonal growth rates are lacking in Saudi Arabia. Despite the collection of information about internal migration in every census and demographic survey, the published data is scarce and of little use to this research. The published data is mostly available in the form of net rates of migration. However, ADA provides some brief, yet useful, information about the population growth of Riyadh city disaggregated by natural and migration growth rates. Additionally, it provides an estimation of annual average migrants to Riyadh. However, there is no information about the migrants' gender, age and marital status or family size. The data provided by ADA is summarised in Table 9.

Table 9 internal migration data original format.

Aspect	Year			
	1987-1990	1990-1996	1996-2004	2004-2016
Population growth	8.8%	8.1%	4.2%	4%
Natural growth	3.3%	3.3%	3%	3.2%
Migration growth	5.5%	4.8%	1.2%	0.8%
Annual avg. of migrants (person)			32500	23000

The format of the available data is conventional for population demographics data. However, data quality represents a formidable obstacle.

B. Data quality evaluation at the city level

The four major factors—contextual, intrinsic, representational, and accessibility—that were previously discussed in 3.2 Data quality dimensions will be followed in the evaluation of data quality.

In terms of contextual DQ, the census data gathering side was carried out on a "De Facto" basis, and then corrected with the addition of those who reside outside the Kingdom of Saudi Arabia (GAS, 1992, 2004, 2010). Additionally, the support of GIS technology was first utilised during the 2004 Census (GAS, 2004). Regarding demographic surveys, the design of GAS surveys is a multi-stage procedure that depends on the geographic context of prior censuses. In addition, the sample design is stratified to take into account the various administrative regions and nationalities. Next, a randomly selected subregion and household sample is used to cover all administrative regions. Targeting families with members with disabilities was one of the goals of the 2001 demographic survey (GAS, 2001). Both census and demographic surveys calculate total fertility rates (TFR) in relation to fertility rates using birth records from the year before the research. Moreover, questions about fertility that are only applicable to women are included in all data collection forms. When calculating death rates, census and demographic surveys use birth records from the year before the research year. In both censuses and demographic surveys, secondary questions concerning the place of birth and current residence are used to gather information about internal migration.

In relation to the element of data format, given the prevalence of demographic information in most data, the existing data is acceptable for the majority of cases, while in a few instances, modest alterations were necessary, such as the aggregation of age groups.

In terms of data timeliness, this is a considerable barrier. This is due to the fact that data collection did not adhere to standard practice for the decennial census. Moreover, the accessible data only dates back 17 years, which is a relatively short time frame that hinders its usability. The absence of the aforementioned components makes it difficult to understand the causes of population growth and validate its patterns. As a result, higher degrees of uncertainty in population predictions are to be expected.

In terms of data collection frequency, the GAS data were gathered seven times over the course of 17 years, from 1999 to 2016. Despite the fact that this is a high frequency rate, the lack of uniformity between time periods hinders the capacity to extrapolate the data.

Almost all population data comes from GAS, which contributes to the consistency of data in classifications and definitions. The reported data is, however, not entirely consistent. For instance, the age groups may be categorised in a compact form in some years, such as age-group (0-4), however in other years, this age-group may be divided into two groups, such as age-group (> 1) and (1-4) years, respectively. A simple solution to this kind of consistency is to aggregate the data. Other sorts of inconsistency, such as for some definitions, cannot be changed in order to account for the variations. For instance, despite the usage of four separate terminologies, such as "number of live births," "number of living children," "number of births," and "number of child survivors," we discover that the majority of definitions are missing from the information processing of birth rates. The lack of a definition for these phrases makes it difficult to determine how they differ and whether their meanings are similar or can be inferred. As a result, it might be nonsensical and pointless for authors to compare these data to one another. These issues necessitate a thorough review of the data labels prior to their application. In our case, only data on the number of live births was employed. Before adopting any data, additional consideration must be given to the intended sample that the surveys use, such as the 2001 demographic survey's focus of individuals with disabilities, which would unintentionally bias the results.

At the Kingdom level, the geographical coverage of the data that is currently available might be deemed sufficient. Moving to smaller geographic regions, however, has an impact on the coverage level. For instance, the population figures for Riyadh city in the 2004 and 2010 censuses are only available in highly aggregated form, and any disaggregated data would necessitate data synthesis. This problem is widespread around the world and is often resolved using different estimating techniques.

Moving on to Intrinsic data quality, we will look into data credibility and error reporting. Given that the information is sourced from government agencies in the Kingdom, it should be highly credible. Even though the data is received from a reliable source, it is not guaranteed to be error-free. The lack of reported error levels in both census and demographic surveys is a drawback of using GAS data, though. There are no sections about level of error or post-enumeration surveys in published census reports. According to the UN (2010), post-enumeration was not part of the 2010 census, and there is no information regarding earlier censuses. Also, neither expected error rates nor confidence levels were mentioned in any of the reports that were made public about demographic surveys.

Given that the data is typically presented as tabulations with disaggregated data, it can be said to have good representational and accessible qualities. Additionally, the availability of data in digital formats has made data processing and manipulation largely simple procedures. Regarding publishing policies, the vast majority of data at the city level and above are accessible online and open to the public. Data at the district level are, however, subject to highly rigorous policies that frequently render them unreachable. By and large, data disclosure controls are considered as a major obstacle to access data.

C. Data preparation at the city level

The three substages of the data preparation stage are data exploration, curation, and future projection. Due to the significance of the data as well as the inconsistencies that were discovered in the data, this stage has been quite laborious. Due to the limited scope of this study, the issues raised were addressed in accordance with their significance and viability. Moreover, rather than promising to correct the data, data preparation should be seen as a start in the correct direction of exploring and correcting the data. Clearly, this chapter and its subsections highlight the variety of data features to take into account, the complexity and disaggregation of the necessary data, and the degree to which subject-matter specialists should be involved in data analysis. Perhaps the largest error could be using the data exactly 'as is'.

C.1 Data exploration at the city level

Four concerning issues were found in the population data. These issues are concerned with: 1) The 2010 census sharp decrease of birth numbers; 2) Unjustifiable fluctuation in some age cohorts at the kingdom level; 3) The sharp decrease in ; and 4) Inconsistent Internal migration rates. The details of these issues are explained in the following sub-sections:

C.1.a The 2010 census sharp decrease of birth numbers

First, the number of births reported in the 2010 census was the lowest among all available data. The rate of decrease cannot be considered as marginal as it averaged 28.5%. See Table 10 for the comparison matrix of birth numbers between year 1999 and 2016.

Table 10 The comparison matrix of birth numbers between year 1999 and 2016.

Year / living new-born		**1999**	**2004**	**2007**	**2010**	**2016**
		568039	**486360**	**527843**	**367598**	**492898**
1999	**568039**	0%	17%	8%	55%	15%
2004	**486360**	-14%	0%	-8%	32%	-1%
2007	**527843**	-7%	9%	0%	44%	7%
2010	**367598**	-35%	-24%	-30%	0%	-25%
2016	**492898**	-13%	1%	-7%	34%	0%

Table 10 shows a comparison matrix between number of live births for the years 1999, 2004, 2007, 2010 and 2016. The comparison is based on the percentage of change between the years. The number of births in year 2010 is 35%, 30% and 25% less than what it used to be in years 1999, 2007 and 2016 respectively. However, the average change ratio when the 2010 births are excluded would range between -13% and 9%. Therefore, the 2010 births numbers are most likely inaccurate. To confirm this issue, other data sources had to be made available. The annual statistical books published by the MOH were used to obtain the number of births in public hospitals at the Kingdom level (See Table 11).

Table 11 The comparison matrix of percentage change among births numbers in public hospitals between 2005 and 2016.

Year/		**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**
live births (1000)		**243.0**	**246.6**	**254.8**	**257.7**	**263.2**	**263.6**	**254.0**	**266.8**	**261.5**	**264.1**	**262.8**	**259.2**
2005	**243.0**	0%	-1%	-5%	-6%	-8%	-8%	-4%	-9%	-7%	-8%	-8%	-6%
2006	**246.6**	1%	0%	-3%	-4%	-6%	-6%	-3%	-8%	-6%	-7%	-6%	-5%
2007	**254.8**	5%	3%	0%	-1%	-3%	-3%	0%	-5%	-3%	-4%	-3%	-2%
2008	**257.7**	6%	4%	1%	0%	-2%	-2%	1%	-3%	-1%	-2%	-2%	-1%
2009	**263.2**	8%	7%	3%	2%	0%	0%	4%	-1%	1%	0%	0%	2%
2010	**263.6**	8%	7%	3%	2%	0%	0%	4%	-1%	1%	0%	0%	2%
2011	**254.0**	5%	3%	0%	-1%	-3%	-4%	0%	-5%	-3%	-4%	-3%	-2%
2012	**266.8**	10%	8%	5%	4%	1%	1%	5%	0%	2%	1%	2%	3%
2013	**261.5**	8%	6%	3%	1%	-1%	-1%	3%	-2%	0%	-1%	-1%	1%
2014	**264.1**	9%	7%	4%	2%	0%	0%	4%	-1%	1%	0%	1%	2%
2015	**262.8**	8%	7%	3%	2%	0%	0%	3%	-1%	1%	-1%	0%	1%
2016	**259.2**	7%	5%	2%	1%	-2%	-2%	2%	-3%	-1%	-2%	-1%	0%

The comparison matrix of percentage change among births numbers in public hospitals is between 2005 and 2016. Unfortunately, the data does not cover births in private hospitals nor is it disaggregated demographically. However, the data covers the period between year 2005 and 2016. Moreover, it could be considered as a large sample to indicate the existence of sharp fluctuations in year 2010. Consequentially, the MOH records indicates a percentage of change of around 10% among the years, which is consistent with the percentage of change when excluding year 2010 in Table 10. Furthermore, the 2010 number of births in public hospitals does not flag any significant decrease when compared to the other years.

C.1.b Unjustifiable fluctuation in some age cohorts at the kingdom level

This issue was discovered while trying to estimate migration flows among the 13 administrative regions in Saudi Arabia. The main aim was to study the regional distribution of Saudi population age-groups between year 2004 and 2010, which could give us some insights about the migration trends and its associated age distribution. Although, the time gap between the years 2004 and 2010 is not an equal 5 years, the comparison between age-groups of the two census data should be relevant to a large extent. The prominent outcome indicated an inflation on most age-groups that is simply not possible to occur. See Table 12 for a comparison between the 2004 census age-groups and its relevant age-groups in year 2010 at the kingdom level and Riyadh region.

Table 12 comparison between the 2004 census age-groups and its relevant age-groups in year 2010 at the kingdom level and Riyadh region.

Shifted age-group by year		Population difference on the geographical level		Growth rate		Crude death rate At KAS level
2004	**2010**	**Al-Riyadh**	**KSA**	**Riyadh / Riyadh (age-group totals)**	**KSA / KSA (age-group totals)**	
0--4	**5--9**	-43151	-237717	-9%	-11%	1% - 2%
5-9	**10-14**	-71	-163690	0%	-7%	0.36%
10--14	**15--19**	16091	-31257	3%	-1%	
15--19	**20--24**	62947	222599	15%	12%	
20--24	**25--29**	107434	421862	30%	27%	
25--29	**30--34**	37282	223484	11%	16%	
30--34	**35--39**	31941	213272	12%	19%	
35--39	**40--44**	-14749	24622	-6%	2%	
40--44	**45--49**	21212	104822	12%	13%	
45--49	**50--54**	29502	67281	22%	11%	

Table 12 shows the age-groups between age (0-4) and (45-49) in year 2004, and the shifted or relevant age-groups that falls between age (5-9) and (50-54) in year 2010. Moreover, the table shows, the change in age-groups in numbers and percentage. Also, it provides the crude death rate at the Kingdom level to anticipate the possible level of change. From Table 12 it can be noted that:

1) The population of Age-group (0-4) in year 2004, which is expected to be in age-group (5-9) in year 2010, decreased by 11% and 9% at the Kingdom level and Riyadh region level respectively. This ratio is about 10 fold the crude death rate which is around 1% to 2%. This decrease is clearly not acceptable from the crude death point of view.

2) Although, all age-groups are expected to decrease when shifted after 5 years given the size of age-groups, most age-groups under study increased remarkably. For example, age groups between age (15 – 34) increased approximately between 10% to 30%. This jump in numbers could not justified for the following reasons:
   a) The Saudi citizenship is strictly granted and is not given to migrants based on how long they stay. Even Saudi mothers cannot give their children the nationality if the father was not Saudi. Therefore, the Saudi population can be considered as a closed system.
   b) Saudi Census adjusts its counting to include those who are living outside the country to or for various purposes. So, the assumption of returning population should be excluded.
   c) The increase ratios are not reasonable and beyond common acceptable levels of error.

Given these above-mentioned reasons, the increase in population numbers among the shifted age-groups cannot be accepted and highlights a major flaw to be addressed in the data. Finally, this problem is not limited to these two censuses, and other issues could be found when comparing the 1999 demographic survey data to the 2004 census or the 2010 census to the 2016 demographic survey. The 2004 and 2010 comparisons were chosen due to being both census data and these highlight more issues surrounding the 2010 data.

C.1.c The sharp decrease in TFR

The rapid and sharp decrease in TFR is a well-known issue in Saudi Arabia. GAS data indicates that TFR plummeted from around 7 births in the 70s and 80s to under 3 births in 2015 with a negative future perspective to be under the replacement rate (2.1 births) by year 2035 (UN, 2019). However, although the author anticipates the decrease in TFR, the magnitude of change seems unrealistic according to his understanding of the Saudi culture and religion drivers. Therefore, this decrease in fertility rate was further investigated from both sides of the literature and available data.

On the literature side, total fertility rates are calculated based on dividing total count of births by the number of all women in childbearing years. Therefore, this rate represents the expected average of the reproduction rate under the assumption that any women who just started her childbearing years will give birth according to the current period trends (Bongaarts and Feeney, 1998; Sobotka and Lutz, 2010). Accordingly, TFR is a synthetic period measure instead of a cohort's births measure (Bongaarts and Feeney, 1998; Ní Bhrolcháin, 2011; Sobotka and Lutz, 2010). Moreover, as a period measure TFR is highly sensitive to changes in timing of childbearing which would inflate or deflate the rate rapidly (Sobotka and Lutz, 2010). Although, TFR is widely used by demographers, it has its known short comings and is not suitable for all purposes, including population forecasts due to its high sensitivity (Ní Bhrolcháin, 2011). Bongaarts and Feeney (1998) claim that TFR could be used to understand the implication of age distribution on the number of births. Alternatives to TFR include Bongaarts and Feeney (1998) adjusted TFR and Butz and Ward (1979) Average Cohort fertility. However, Ní Bhrolcháin (2011) argue that there is no empirical evidence to justify the use of these indices. Moreover, Rallu and Toulemon (1994) believe that these synthesised indices which include some degrees of estimation are in a similar position as TFR because they could be biased in terms of level and trends. On the other hand, the actual number of births of women is represented by the Completed Fertility Rate (CFR). According to Bongaarts and Feeney (1998), CFR is a true measure of reproduction that is representing the average count of births by women who finished their childbearing years. Moreover, CFR is a stable measure, unlike TFR which is subject to period changes (Schoen, 2004). However, its disadvantage lies in reflecting past reproductive experiences dating back to two to three decades ago (Bongaarts and Feeney, 1998). Schoen (2004) states that:

> *"One can readily concede that period conditions can influence both the timing and the level of cohort reproduction and still view completed cohort fertility as the most informative measure of fertility behavior. Moreover, early cohort goals can be important even if they later change." (Schoen, 2004, p. 802)*

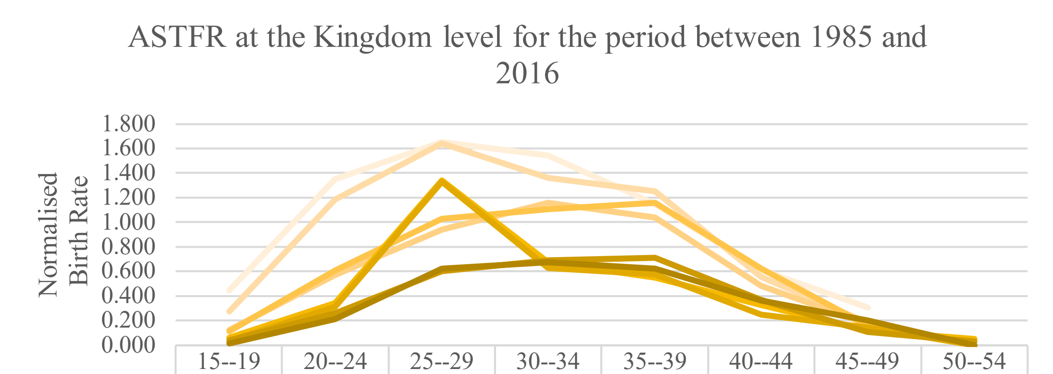
Thus, he promotes the juxtaposition of the use of both indicators to advance adjusted TFR measurements to reflect actual CFR.

On the practical side, both TFR and CFR fertility measures were investigated. CFR were obtained and compared to TFR for the period between 1975 and 2015. See Table 13 for the comparison of CFR and TFR for the period between 1975 and 2015.

Table 13 Saudi Arabia CFR and TFR for the period between 1975 and 2045.

Childbearing end year (reference year)	**Childbearing start year**	**TFR**	**CFR**
2015	**1980**	2.72	4.49
2010	**1975**	3.22	5.16
2005	**1970**	3.65	5.46
2000	**1965**	4.40	5.54
1995	**1960**	5.55	5.40
1990	**1955**	6.22	5.54
1985	**1950**	7.02	5.42
1980	**1945**	7.28	5.11
1975	**1940**	7.30	4.73

In Table 13 the childbearing end-year was considered as the reference year. Moreover, the start of childbearing years was calculated based on a maximum of 35 years of childbearing period. Furthermore, the CFR values are averaged based on all available CFR data in (2004, 2007, 2016). Table 13 shows that despite the rapid decrease in TFR, the CFR remained stable for most of the time at around 5.2 births. Although a monotonic decrease can be observed in CFR that has started from year 2000, this monotonic decrease caused the CFR to decrease by 1 birth for the past 15 years compared to 2 births in TFR for the same period. Strong evidence on how the TFR can be misleading is found when we compare the women cohort TFR, for the reference years 1975 and 1980, with their corresponding CFR in reference years 2010 and 2015. In other words, we compare the estimated TFR for women who would start their childbearing in year 1975 and 1980 with the actual numbers of average births (CFR) for those who started in their childbearing years in 1975 and 1980 (referenced as year 2010 and 2015). By this comparison we can directly compare the TFR with CFR for two women cohorts (1975 and 1980). Consequently, while the TFR for 1975 and 1980 were at 7.3 and 7.28 respectively, the CFR for these cohorts were at 5.16 and 4.49 births respectively. The difference between TFR and CFR accounts for an average around 2.5 births per women, that is about 51% increase in the TFR estimates. Therefore, the author believes that change in the TFR is due to timing effects more than an actual severe drop in fertility. To understand the changes of period trends we will compare all available TFR from year 1985 to 2016. See Figure 13 AS-TFR at the Kingdom level for the period between 1985 and 2016.



	15--19	20--24	25--29	30--34	35--39	40--44	45--49	50--54
1985	0.446	1.351	1.654	1.543	1.146	0.610	0.306	
1987	0.273	1.182	1.642	1.362	1.253	0.556	0.184	
1999	0.128	0.572	0.939	1.160	1.037	0.482	0.168	
2001	0.115	0.607	1.027	1.104	1.161	0.621	0.162	0.000
2004	0.062	0.339	1.340	0.673	0.547	0.327	0.120	0.051
2007	0.037	0.312	1.333	0.627	0.577	0.250	0.142	0.000
2010	0.039	0.256	0.599	0.691	0.712	0.368	0.106	0.031
2016	0.016	0.213	0.622	0.677	0.620	0.362	0.202	0.000

Figure 13 AS-TFR at the Kingdom level for the period between 1985 and 2016.

Figure 13 shows all values of AS-TFR that are computed over the period between 1985 and 2016. By and large, the values of TFR over this period can be characterised by their decrease over time and peak shift to elder age-groups. These two characteristics can be detailed in four patterns as follows:

The first pattern occurred in the period (1985-1987) with TFR of approximately 7 births. The peak of births belongs to age-group (24-29) which is a little above 1.6 births. The second pattern was observed between (1999 - 2001). TFR values decreased to about 4.5 births and, the peak of births moved towards a larger and wider age-group (30 - 39). The peak births ratio lies between 1 to 1.2 births. The third pattern happened over the period (2004-2007). The further decrease in TFR is clear and accounts for less than 3.5 births. However, strikingly, the peak in births shifted backward to age group (25-29) and were at a higher level than all age groups in the second pattern at an average of 1.3 births. This fluctuation is a clear indication of the high sensitivity of TFR as a measure. Moreover, such changes raise questions about the influences that could led this fluctuation. The fourth pattern was for the period (2010-2015), where birth rates reach less than the level of 3 births. It also shows that the peak of births moved again to include the age groups between (30 - 39) and those that fall below 0.8 births. These four patterns clearly show the presence of a gradual decrease in TFR, as well as a change in their peak reproductive age groups.

Before moving to another element in the population, it is important to clarify that all numbers belong exclusively to the Saudi population. The UN (2019) data is comparable to the GAS data. Moreover, the birth ratios of all-women and ever-married women were tested and found only marginally different.

C.1.d Inconsistent Internal migration rates

Given the scarcity of fertility rates in Saudi Arabia, we compared the shifted age-groups of 2004 censuses to its relevant age-groups in 2010 census. This comparison was done for all 13 administrative regions of Saudi Arabia. The result of this comparison highlighted the unjustifiable patterns of change in the population numbers that was first reported in the section C.1.b Unjustifiable fluctuation in some age cohorts at the kingdom level. See Table 14 the percentage of change between some shifted age-groups of year 2004 into year 2010.

Table 14 the percentage of change between some shifted age-groups of year 2004 into year 2010.

Shifted Age-groups		Percentage of change													
2004	**2010**	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**	**R11**	**R12**	**R13**	**KSA total**
0--4	**5--9**	-23%	-27%	-30%	-28%	-19%	-25%	-35%	-26%	-19%	-27%	-28%	-35%	-19%	-11%
5--9	**10--14**	0%	-8%	-12%	-5%	-7%	-8%	-23%	-5%	-14%	-12%	-13%	-16%	-6%	-7%
10--14	**15--19**	3%	4%	0%	2%	-7%	-6%	-1%	-2%	-10%	-9%	-3%	-11%	-9%	-1%
15--19	**20--24**	15%	14%	7%	10%	19%	2%	24%	3%	10%	-1%	13%	-1%	7%	12%
20--24	**25--29**	30%	26%	25%	23%	30%	23%	29%	19%	21%	29%	35%	27%	20%	27%
25--29	**30--34**	11%	19%	15%	19%	9%	18%	4%	22%	20%	40%	9%	28%	15%	16%
30--34	**35--39**	12%	28%	16%	19%	9%	23%	11%	25%	15%	40%	17%	29%	11%	19%
35--39	**40--44**	-6%	4%	9%	12%	-9%	16%	7%	12%	9%	18%	15%	21%	-1%	2%
40--44	**45--49**	12%	12%	15%	19%	5%	24%	25%	14%	18%	16%	27%	26%	12%	13%
45--49	**50--54**	22%	8%	5%	8%	13%	13%	21%	0%	8%	-3%	10%	2%	0%	11%

R1: Riyadh region
R2 – R13: all other regions

Table 14 shows that there are obvious unjustifiable patterns of change in the population numbers across the 13 regions of the Kingdom. For example, there is a population reduction in age-groups (0-4) and (5-9) in all regions, which accounted for a national decrease at 11% and 7% respectively. This decrease is far beyond any natural crude mortality rate. Moreover, for the ages between 15 and 49 there is an inflation pattern that is also not possible to have occurred naturally. These patterns of change made the study of migration flows impossible using GAS data. Alternatively, the author cn use the net migration rates in Riyadh city estimated by the ADA demographic surveys.

C.1.e Final remarks on identified data issues

Most of the above-mentioned problems would affect the credibility of GAS data and any outcomes of using its data. However, the author acknowledges his limited ability to clear all the doubts about the issues that have been spotted, due to the limited availability of data and the adoption of approximation techniques to overcome lack of data. The approximation of data, to some extent, would create inexistent error or inflate the margins of error. Overall, homogeneous datasets have been one of the main challenges in urban simulation if not the main challenge. The burden to overcome data issues lies on the researchers and their resources to find alternative datasets, curating the data when possible and estimating missing data.

C.2 Data curation methods

Due to the interrelated association between population numbers and its growth factors (fertility, mortality and immigration), the author believes that it is not possible to correct any encountered issues in separation from all other factors of population change. For example, adjusting fertility rates to a higher level would directly increase the number of births and new-born deaths simultaneously. Therefore, to account for these interrelations of change, a CCM method has been used as a control mechanism to simulate population growth and to correct the suspected components in population numbers. This CCM has been manually coded in AnyLogic (as an accountant model) to run under its optimisation engine for the data curation process. The reason behind adapting a calibration procedure is to vary the population growth parameters to achieve a set of predefined numbers of population, as objectives to the calibration process that is believed to be more accurate. These objectives consist of numbers of enrolled students in elementary and secondary schools and selected population age-groups from the 2004 census data. The numbers of enrolled students are believed to provide a sound reference point for correcting the population data due to being collected from the annual students' records. It is worth noting that the main aim of this model is to approximate the population numbers to the best-known external references available within previously observed limits of the drivers of change. This approximation stance inherently acknowledges that external data has short-comings in terms of availability and accuracy. However, this approach is indispensable for our purposes. Figure 14 shows the curation model framework for the population data between the years 1999 and 2014 with this its objectives and inputs.

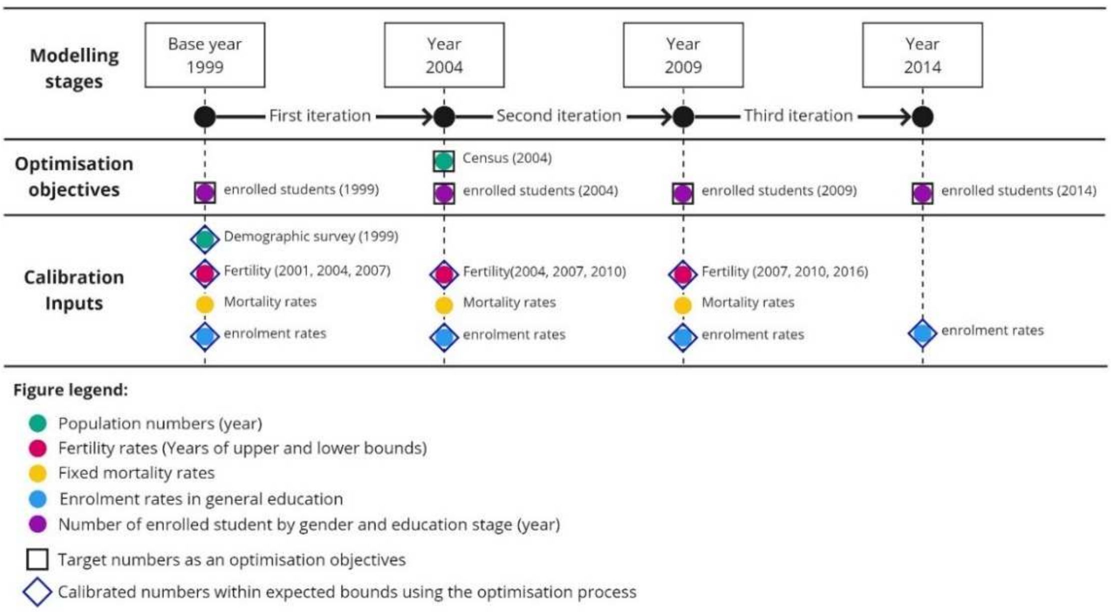


Figure 14 The curation model framework for the population data between year 1999 and 2014.

Figure 14 shows that the curation model consists of four modelling stages (1999, 2004, 2009 and 2014) and two types of inputs. The two types of inputs are calibration objectives and constraints. The calibration objectives are based on external data references such as the numbers of enrolled students and census data. The calibration side contains mainly the CCM inputs such as the base year population, fertility rates and mortality rates with the addition of enrolment ratios of schooling age population in elementary and secondary schools. The following subsections will discuss the model inputs, methodology, and results.

C.2.a The inputs of data curation process:

For the calibration objectives, the number of enrolled students and the 2004 census data were used. The available records of enrolled students cover the period between 1999 and 2018. Moreover, these numbers belong to registered students in day education and do not include night school, illiterate and special needs education. Also, the students' records are disaggregated by educational stages and gender. However, the data is lacking for the years 2008 and 2009. Moreover, the reported numbers for the year 2010 are inconsistent. For example, enrolled girls in the year 2010 are significantly lower than year 2011 for middle and secondary schools by 74% and 37% respectively. Meanwhile, average percentage of change between all consecutive years excluding year 2010 is 1.5% with a standard deviation of ± 6.0%. To overcome the problem of fewer students in year 2010 and any other unknown fluctuations, we will average the number of students for the wanted years (1999, 2004, 2009 and 2014) with two surrounding years as possible. Consequentially, for the year 1999, the averaged values belong to the years 1999, 2000 and 2001. This is because no records are available before year 1999. Then, for the year 2004, the values of years 2003, 2004 and 2005 were averaged. After that, for the year 2009, due to the lack of records for year 2009, the fluctuations appear in the 2010 girls' records. The extent of averaged years was relaxed as follows. The averaged data for boys belongs to 2007, 2009, 2011 and 2012, and for the girls belongs to 2007, 2011 and 2012. Finally, for the 2014, year the averaged values belong to 2013, 2014 and 2015. See Table 15 for the averaging years, values and percentage of change for the numbers of enrolled students.

Table 15 The averaging years, values and percentage of change for the numbers of enrolled students used in the curation model inputs.

The targeted years	**1999**	**2004**	**2009**	**2014**
Averaged years for boys data	**(99,00,01)**	**(03,04,05)**	**(07,10,11,12)**	**(13,14,15)**
Elementary boys	1188503	1245237 (+5%)	1272086 (+2%)	1318663 (+4%)
Middle boys	578381	584956 (+1%)	600508 (+3%)	616586 (+3%)
Secondary boys	387720	474367 (+18%)	539924 (+12%)	581057 (+7%)
Averaged years for girls data	(99,00,01)	(03,04,05)	(07,11,12)	(13,14,15)
Elementary Girls	1096040	1131617 (+3%)	1206334 (+6%)	1301839(+7%)
Middle Girls	485777	519200 (+6%)	560439 (+7%)	606098(+8%)
Secondary Girls	363668	446578 (+19%)	506363 (+12%)	565450 (+10%)

(%) percentage of change from last five-years.

Table 15 shows the range of years used to produce the average number of enrolled students as well as, the percentage of change from the past five years. The average percentage of change for all values is 7% with a standard deviation of ± 5%. The significant observations from Table 15 can be concluded with respect to two points. First, the percentage of change in secondary stage is higher than the standard deviation limits. Moreover, the percentage of change in girl’s education is approximately 3% higher than the percentage of boy’s education. While the average change for boys is about 6%. The average percentage of change for girl’s education is roughly 9%. These higher ratios of change might point at improvements in levels of the enrolment ratio in school age population at the Kingdom level.

Another data that is used as a reference data is the 2004 census. The reason behind including the 2004 census data is to add another layer of credibility to the calibration results by replicating population enumeration data. This data can be of great controlling influence if considered exclusively as a calibration objective for all age-groups. Otherwise, it can be partially considered in many ways if there are overlapping objectives and there is a need to relax the calibration process. For example, choosing the census population aged (0-4) as a calibration objective will force the calibration of birth rates to be within a stricter solution space. Another example is that for the schooling ages between 6 and 18 years old, there would be overlapping objectives when considering the number of enrolled students besides census data of similar age-groups. Therefore, the inclusion of census data is subject to trail-and error in the optimisation process, that could be judged with the outcome’s level of error.

- Base year population

In terms of the base year population, the 1999 demographic survey data has been chosen as the starting point for two reasons. First, 1999 is the earliest comprehensively available dataset. Second, the period between the 1999 demographic survey and the 2004 census is five years, which enables a seamless shifting of age-groups between these two reference years. However, due to the general absence of error level reporting in GAS data, the 1999 numbers are based on demographic surveying instead of population enumeration. Therefore, the 1999 dataset data cannot be considered error free and, it is likely to have a higher margin of error than previously investigated data census. Smith, Tayman and Swanson (2013) confirm this higher possibility of error by stating that that in the US, postcensal population estimation are subject to higher levels of error than census which could be considered as best practice. However, the possible extent of the margin of error for the 1999 dataset is unknown to the author. To limit the guesstimate margin of error, the author uses analogies to several simple extrapolation methods to forecast the population of three counties in Florida, USA in year 1990 and using data from 1980, Smith and Shahidullah (1995) concluded that the mean absolute percentage of error for forecasting individual age-groups can range from 20% to 29%. Therefore, under the assumption of the worst-case scenario, the calibration limits were relaxed for the base year population to a maximum of +/- 25%. See

Table 16 for the calibration limits for the 1999 dataset population data. Consequently, this approach will accept a restricted possibility of error in the base year population to assist in achieving the optimisation targets more accurately.

Table 16 the calibration limits for the 1999 dataset population data.

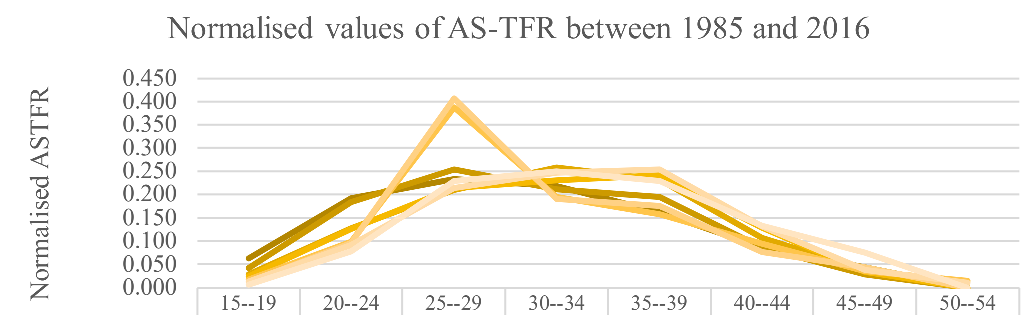
age-	Female			Male		
group	**actual**	**(-25%)**	**(+25%)**	**actual**	**(-25%)**	**(+25%)**
0--4	1296673	1006928	1678214	1342571	972505	1620841
5--9	1123187	872671	1454451	1163561	842390	1403984
10--14	961877	741050	1235084	988067	721408	1202346
15--19	816754	606190	1010316	808253	612566	1020943
20--24	686072	448783	747971	598377	514554	857590
25--29	552053	390779	651298	521038	414040	690066
30--34	442178	322814	538024	430419	331634	552723
35--39	369260	273918	456530	365224	276945	461575
40--44	290278	216575	360958	288766	217709	362848
45--49	219356	161177	268629	214903	164517	274195
50--54	173671	123892	206486	165189	130253	217089
55--59	141947	96028	160046	128037	106460	177434
60--64	112778	75738	126230	100984	84584	140973
65--69	91576	95638	159396	127517	68682	114470
70--74	55014	65721	109535	87628	41261	68768
75--79	32960	41741	69568	55654	24720	41200
80+	49820	53372	88953	71162	37365	62275

- Fertility rates

Moving into aspects of the fertility rates in the calibration inputs, given the problem of sharp decline in TFR as discussed in page 94, the author proposes the use of CFR instead of TFR. However, CFR are not ready for use in their original form and will be estimated as Age Specific Fertility Rates (AS-FR) using data synthesis and estimation as part of the calibration process. In other words, the AS-FR will be estimated based on CFR during the calibration process of population numbers. Moreover, the AS-FR synthesis will be constrained by ranges that are a product of synthesising normalised and averaged AS-TFR with disaggregated CFR by gender and year. The estimation process of AS-FR based on the CFR was done though the following steps:

1) Normalisation of AS-TFR data.
   Age-Specific Total Fertility Rates are available for the years 1985, 1987, 1999, 2001, 2004, 2007 and 2010. However, all data before year 2001 are not disaggregated by gender. Therefore, gender aggregated ratios were used to generate a single normalised version of AS-TFR. Moreover, the rates used in this study belongs to the ever-married women instead of all women. This is because the estimation ranges which are based on the variation of observed data are larger than the differences between the ever-married and all-women total ratios. The difference between ever-married and all-women CFR ratios are marginal at 0.9%, 1.3% and 1,8% for the years 2000, 2005 and 2010 respectively. It is noted that there is an increasing gap between the two ratios which can be incorporated in future studies. Moreover, the households-based city model will require ever-married women biased AS-FRs due to the use of marital status as a controlling rule for births to occur. See Figure 15 for Normalised AS-TFR at the kingdom level for the period between 1985 and 2016.

   Figure 15 shows the magnitude of births available as AS-TFR. Two remarkable features in the data can be identified first, the shift in peak reproduction age-groups from age-group (25-29) towards age-group (35-39), and second, the sudden jump in fertility for the age-group (25-29) in years 2004 and 2007. The cause of this jump in such a significant level is unknown to the author and beyond the scoop of this research. Therefore, a new step was suggested to deal with this surprising fluctuation.



	15--19	20--24	25--29	30--34	35--39	40--44	45--49	50--54
year 1985	0.063	0.191	0.234	0.219	0.162	0.086	0.043	0.000
year 1987	0.042	0.183	0.254	0.211	0.194	0.086	0.029	0.000
year 1999	0.028	0.127	0.209	0.259	0.231	0.107	0.038	0.000
year 2001	0.024	0.127	0.214	0.230	0.242	0.129	0.034	0.000
year 2004	0.018	0.098	0.387	0.195	0.158	0.095	0.035	0.015
year 2007	0.011	0.095	0.407	0.191	0.176	0.076	0.043	0.000
year 2010	0.014	0.091	0.214	0.247	0.254	0.131	0.038	0.011
year 2016	0.006	0.079	0.229	0.250	0.229	0.134	0.075	0.000

Figure 15 Normalised AS-TFR at the kingdom level for the period between 1985 and 2016.

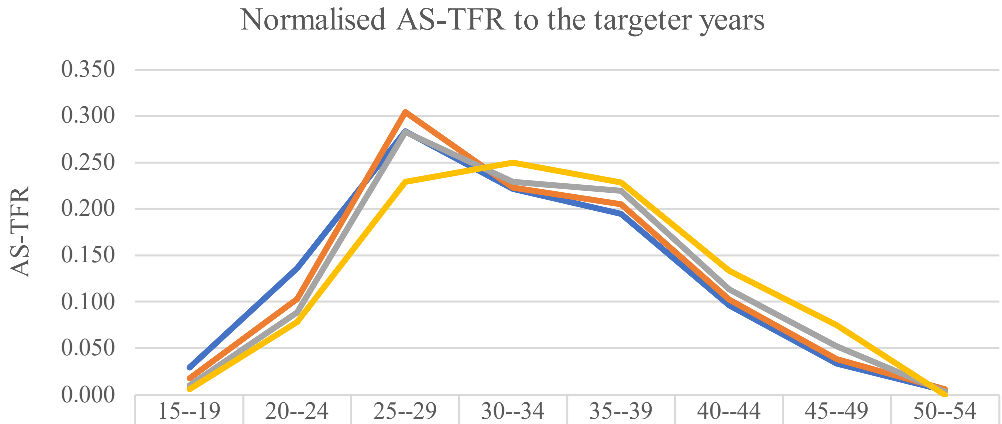
2) Averaging the normalised rates of comparable years.

The reasoning behind averaging the data is driven by three aspects; first, to create a homogenous transition of normalised AS-TFR between the consequent years; second, to overcome the lack of AS-TFR observations for the targeted years in the calibration process of fertility rates; third, to define the calibration limits of AS-FRs, for the curation model, based on the selected averaging data. Several averaging combinations were tested to create normalised AS-FR for the modelled years 1999,2004, 2009 and 2016. The combination of years used for creating the normalised averages and their calibration limits were as follows:

1) Year 1999: (1987, 1999 and 2004)
2) Year 2004: (1999, 2004, 2007 and 2010)
3) Year 2009: (2007, 2010 and 2016)
4) Year 2014: (2016)

These averaging years were chosen in arbitrary manner due to having a very limited number of observations to cancel the 2004 and 2007 fluctuations in peak reproduction ages. Moreover, the general trend had to reflect the decrease in fertility with the shift of reproduction peak to older age-groups which was not attainable using a systematic averaging of available years. Therefore, for the year

2004, the averaging extent had to be wider to smooth the data as much as possible. Moreover, for year 2014, the 2016 rates solely were used for being the most reflective of the current trends. See Figure 16 for the averaging of normalised AS-TFR to the targeted years.



	15--19	20--24	25--29	30--34	35--39	40--44	45--49	50--54
1999	0.030	0.136	0.284	0.221	0.195	0.096	0.034	0.005
2004	0.018	0.103	0.304	0.223	0.205	0.102	0.038	0.006
2009	0.010	0.088	0.283	0.229	0.220	0.114	0.052	0.004
2014	0.006	0.079	0.229	0.250	0.229	0.134	0.075	0.000

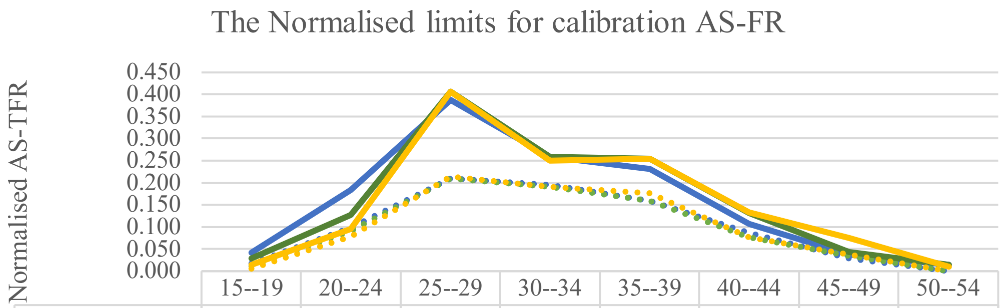
Figure 16 The averaging of normalised AS-TFR to the targeted years.

The main features from the averaging process shown in Figure 16 can be summarised in four points:

1) There is a monotonic decrease in fertility rates for the age-groups between 15 and 24.
2) For the age-group (25-29) the smoothing result caused the 1999 rate to increase about 30% from 0.209 to 0.298 births. However, from the year 2004 rate decreased about 24% from 0.407 to 0.308 births. Ideally, we wanted the rate of year 2004 to be lower than 1999 which was not possible despite using a wider range in averaging. Nevertheless, the fertility trend is generally decreasing.
3) It can be noted that there is compensating shift for the declining fertility rates in ages younger than 30 by fertility increases in all ages between 30 and 50.
4) The rates of age- group (50-55) are largely insignificant but could described as decreasing.

It can be argued that due to the fluctuation caused by the 2004 and 2007, these data can be removed. The author believes that these variations should be accepted as a natural part of TFR sensitivity as indicator. Therefore, the averaging approach should be sufficient to include and smooth all differences for calibration purposes.

Moving to creating the calibration limits for the AS-FR and using the same normalised values in the averaging process, the min and max values within each averaging combination were extracted. This approach was preferred over setting a constant margin of error for each AS-FR, because the calibration process was not intended to be deterministic in following the distribution of averaged values. However, the calibration variation should be within observed ratios. See Figure 17 for normalised limits for calibrating AS-FRs for years 1999, 2004 and 2009.



	15--19	20--24	25--29	30--34	35--39	40--44	45--49	50--54
1999 min	0.018	0.098	0.209	0.195	0.158	0.086	0.029	0.000
1999 max	0.042	0.183	0.387	0.259	0.231	0.107	0.038	0.015
2004 min	0.011	0.091	0.209	0.191	0.158	0.076	0.035	0.000
2004 max	0.028	0.127	0.407	0.259	0.254	0.131	0.043	0.015
2009 min	0.006	0.079	0.214	0.191	0.176	0.076	0.038	0.000
2009 max	0.014	0.095	0.407	0.250	0.254	0.134	0.075	0.011

Figure 17 The normalised limits for calibrating AS-FRs for years 1999,2004 and 2009.

Figure 17 shows that the calibration limits for years 1999 and 2004 are very similar. Moreover, the lower bounds have very little variation. Furthermore, the upper limits of age-groups below age 25 have a decreasing pattern. The age-groups beyond age 34 have an increasing pattern, except that of age-group (50-54). These limits are mainly created for the calibration to be within known boundaries.

3) First generate AS-FR by synthesising CFR and normalised AS-TFR:

   because normalised AS-TFR was not disaggregated by birth's gender, due to the absence of gender in AS-TFR data before year 2001. The synthesis of AS-CFR was disaggregated by the gender-specific CFR while the CFR of male births in 1999, 2004 and 2009 were at 2.788, 2.853 and 2.689 respectively. The CFR of female births were at 2.709, 2.680 and 2.565 respectively. Consequently, the total CFR for all births were at 5.586, 5.533 and 5.254 respectively. Second, the gender-specific CFR were multiplied by the normalised calibration limits of AS-TFR generated in the previous step. See Table 17 for the calibration limits of AS-FRs based on synthesising CFR with normalised AS-TFR for male births. Also Table 18 shows the calibration limits of AS-FRs based on synthesising CFR with normalised AS-TFR for female births.

Table 17 The calibration limits of AS-FRs based on synthesising CFR with normalised AS-TFR for male births.

AS-FR	**1999**		**2004**		**2009**	
year	**min**	**max**	**min**	**max**	**min**	**max**
15--19	0.051	0.122	0.032	0.081	0.016	0.037
20--24	0.282	0.527	0.261	0.364	0.211	0.256
25--29	0.602	1.114	0.597	1.160	0.575	1.094
30--34	0.560	0.744	0.546	0.738	0.514	0.672
35--39	0.455	0.665	0.451	0.725	0.474	0.683
40--44	0.248	0.309	0.217	0.374	0.205	0.359
45--49	0.082	0.108	0.099	0.123	0.102	0.200
50--54	0.000	0.042	0.000	0.042	0.000	0.030
Total	2.281	3.632	2.204	3.607	2.097	3.330

Table 18 The calibration limits of AS-FRs based on synthesising CFR with normalised AS-TFR for female births.

AS-FR	**1999**		**2004**		**2009**	
year	min	max	min	max	min	max
15--19	0.048	0.115	0.030	0.076	0.015	0.035
20--24	0.265	0.496	0.245	0.342	0.201	0.244
25--29	0.567	1.049	0.561	1.090	0.549	1.043
30--34	0.527	0.700	0.513	0.693	0.491	0.641
35--39	0.429	0.626	0.424	0.681	0.452	0.652
40--44	0.234	0.291	0.204	0.352	0.195	0.342
45--49	0.077	0.102	0.093	0.116	0.097	0.191
50--54	0.000	0.040	0.000	0.039	0.000	0.028
total	2.147	3.419	2.071	3.389	2.001	3.177

It is worth noting that the limits for calibrating AS-FR based on the CFR can range from around 4 to 7 births.

The above-mentioned steps were created to avoid the use of the non-suitable AS-TFR measurement by synthesising a cohort-based Age-specific fertility rates for both births' genders. At this stage, the synthesised limits for estimating the age-specific cohort fertility rates for the years 1999, 2004 and 2014 are ready as inputs for the curation model. The calibration process in the calibration model will estimate the values that should replicate the numbers of enrolled students and the selected age-group from the 2004 census.

- Mortality rates

In the model, survival rates are utilised to account for death rates. The ratio utilised is the average of data from the years 2001, 2004, 2007, 2010, and 2016. The inclusion of 2001 despite its bias towards the impaired population was due to the lack of earlier data to replace 1999. The survival data are disaggregated by age and gender and show no significant variations among them. See Table 19 for averaged age-specific survival rates at the kingdom level disaggregated by gender.

Table 19 Averaged age-specific survival rates at the kingdom level disaggregated by gender.

Age group	**Survival rate**	
	Male	**female**
0--4	0.981	0.978
5--9	0.996	0.997
10--14	0.996	0.997
15--19	0.991	0.998
20--24	0.988	0.997
25--29	0.991	0.997
30--34	0.994	0.998
35--39	0.993	0.996
40--44	0.987	0.991
45--49	0.988	0.991
50--54	0.977	0.983
55--59	0.964	0.970
60--64	0.924	0.956
65--69	0.906	0.954
70--74	0.717	0.846
75--79	0.799	0.854
80+	0.520	0.596

- Enrolment ratio in general educational stages in Saudi Arabia

Given that the number of students enrolled in general education does not fully represent the school-age population due to factors such as living in rural areas with no nearby schools or dropping out of school for personal reasons, it became vital to take enrolment ratio in general education into account. Therefore, we must transform the model generated school-age population into numbers of enrolled students. The conversion of school-age population numbers into net enrolled students will allow us to evaluate the calibration results by comparing the actual and estimated net enrolled student numbers. Most net enrolment ratio data comes from MOE and UNICEF statistics on childhood. However, these reports only consider elementary and secondary school enrolment ratios. See Table 20 for net enrolment ratios at the kingdom level for elementary and secondary education by source year and gender.

Table 20 Net enrolment ratios at the kingdom level for elementary and secondary education by source year and gender.

Stage	Source	Year	Male	Female	Average
Elementary	(MOE, 2018)	2017	96.48		96.48
	(UNICEF, 2011)	2011-2016	97	98	97.5
	(UNICEF, 2009)	2005-2009	85	84	84.5
	(MOE, 2000)	1999	92.9		92.9
	(UNICEF, 2002)	1995-1999	60	56	58
	(UNICEF, 2000)	1990-1996	63	60	61.5
	(MOE, 2000)	1990	86		86
Secondary	(MOE, 2018)	2017	73.19		73.19
	(UNICEF, 2011)	2011-2016	76	70	73
	(UNICEF, 2002)	1995-1999	72	65	68.5
	(UNICEF, 2000)	1990-1996	65	57	61

A few points can be highlighted by examining the ratios presented in Table 20:

1) By and large there is an increasing trend in the enrolment ratios in elementary and secondary education.
2) The ratios reported by the (UNICEF, 2000, 2002) for the elementary education are considerably lower for the period between 1990 and 1999 when compared to the (MOE, 2000) ratios.
3) For the period between 1990 and 1999, the elementary ratios reported by UNICEF are lower than enrolment ratios of the secondary stage. This difference is unexpected as normally elementary net enrolment ratios are higher than secondary net enrolment ratios. Consequentially, the UNICEF data for the period before year 2000 might not be accurate.

One could suggest the exclusion of the UNICEF elementary ratios, given the availability of the MOE 1990 ratios as a starting reference point. On the other hand, the use of wide band ratios to set the calibration limits of net enrolled students would be useful in relaxing the calibration process. The author believes that the UNICEF ratios for elementary education are exceptionally low to be included. Moreover, given that the ratios are following an increasing trend, it would be reasonable to set the calibration limits for the enrolment ratios with increasing bands for the targeted years 1999, 2004, 2009 and 2014. In other words, the calibration of net enrolment ratios should not be affected by the UNICEF low ratios but must reflect the improvement in net enrolment ratios. To understand the net enrolment data in detail and how the calibration limits will be prepared, see Figure 18 for the elementary stage enrolment ratios at the kingdom level for the period between 1990 and 2017 and for the Enrolment ratios of elementary schools in Saudi between 1990 and 2017. Moreover, Figure 19 shows the secondary stage enrolment ratios at the kingdom level for the period between 1990 and 2017.

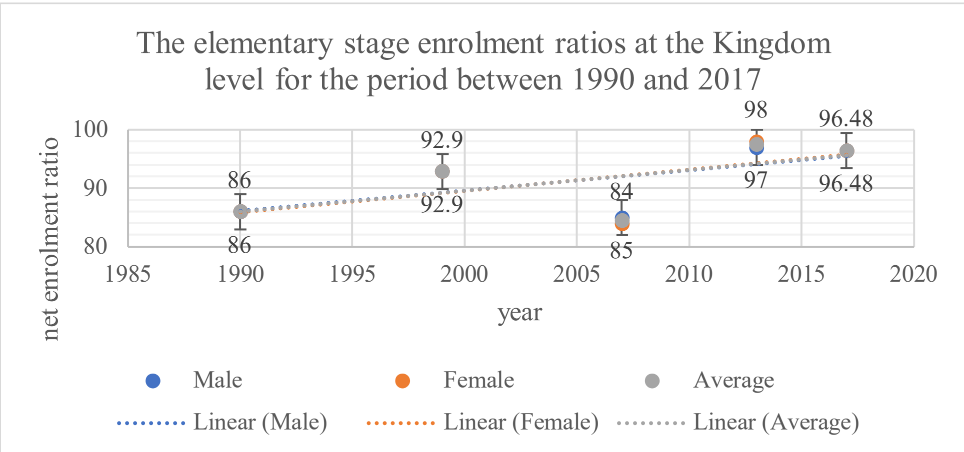


Figure 18 The elementary stage enrolment ratios at the kingdom level for the period between 1990 and 2017.

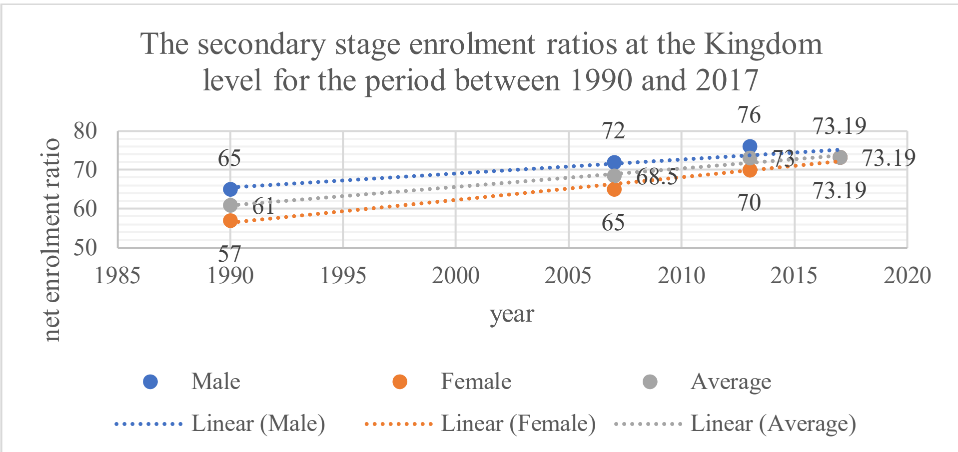


Figure 19 The secondary stage enrolment ratios at the kingdom level for the period between 1990 and 2017.

For elementary education, Figure 18 shows that the enrolment ratios are almost identical for male and females. This is mainly because MOE reports net enrolment ratios aggregated for both genders. Moreover, the ratios reported by UNICEF indicates about a 1% gap between both genders. The average enrolment ratio was around 90% in year 1999 and reached around 96% in 2014. Furthermore, the minimum and maximum values range between 86% and 98%. On the other side, for the secondary education, Figure 19 shows that there is an evident gap between net enrolment levels of male and female education. The gap between the two genders was about 8% in 1990, then it decreased to 6% in 2014. The average enrolment ratio for both genders was around 65% in 1999 and above 73% in 2014. Also, minimum and maximum values range between 61% and 73.19%. For the purpose of creating calibration limits of net enrolment ratios, the following assumptions were considered:

1) Given that the observations are dispersed in time between 1990 and 2017, they have been reorganised into 5 year terms beginning with 1990 and ending with 2015.
2) Considering the upward trend in net enrolment ratios, a constant ratio was proposed for the enhancement of the net enrolment ratio. This ratio was computed using the difference between the minimum and maximum values observed between 1990 and 2015.
3) Given that the gender gap in elementary education has about 1%, the constant ratio of improvement in net enrolment was aggregated for both genders.
4) Given that the gender gap in secondary education averaged around 7.5%, the constant ratio of improvement in net enrolment was disaggregated by gender.
5) The maximum enrolment rate was limited to 99.9%.

See Table 21 for the limits for calibrating net enrolment ratios at the Kingdom level disaggregated by education stage and gender for the period between 1999 and 2014.

Table 21 The limits for calibrating net enrolment ratios at the Kingdom level disaggregated by education stage and gender for the period between 1999 and 2014.

Education stage	**Elementary stage**		**Secondary stage**			
Gender	**both genders**		**male**		**female**	
Constant ratio	2.4%		2.2%		3.2%	
Calibration limit	**min**	**max**	**min**	**max**	**min**	**max**
year 1999	90.8%	93.2%	69.8%	72.2%	63.5%	66.7%
year 2004	93.2%	95.6%	72.2%	74.6%	66.7%	70.0%
year 2009	95.6%	98.0%	74.6%	77.0%	70.0%	73.2%
year 2014	98.0%	99.9%	77.0%	79.4%	73.25	76.4%

Table 21 shows that all the net enrolment rates were considered from the base year of the curation model which is the year 1999. Furthermore, the improvement in elementary education would reach more that 100% if it remained at the same ratio. Therefore, the ratio was set to near 100%. Also, the constant ratio of improvement for the girl's secondary education was higher than secondary male education and elementary education, which could be due to the Kingdom's efforts in closing the gender gap in education.

C.2.b Final remarks on the curation model inputs

The data curation model should be able to use the calibration objectives and inputs that have been completed by this point. Detailed explanations of the available data, significant aspects of the values, data smoothing, and methodology for determining calibration limits for all desired data were provided by the author. Despite the author's efforts to narrow the calibration limitations, he feels that there is still a need for improvement. More subject matter expertise in migration, and education etc. with more access to data sources would be ideal for this.

C.3 The data curation model

As mentioned earlier, the population data correction process uses CCM as a core method and iterates through four steps. These steps represent the years 1999, 2004, 2009 and 2014. Each step has its unique calibrated inputs, fixed inputs and predefined optimisation objectives. In total 96 unique parameters are calibrated for the model. The main objective of this process is to approximate the population numbers and its change parameters for the targeted years by reproducing the numbers of registered students in elementary and secondary schools. This goal is achieved by calibrating the 1999 base year population, the population reproduction AS-FR, and the net enrolment level for the two general education stages. The only fixed input in this model is age-specific morality rates.

C.3.a Process workflow

For the first iteration, the model seeks to calibrate the population numbers for the year 1999 by age-group and gender and, the number of enrolled students by education stage and gender. The former is essential as being the base to all subsequent estimated population numbers. The latter is considered as the main reference to evaluate the goodness of fit of the estimation. Then, for the second iteration, the model objectives include the number of enrolled students by education stage and gender and some selected age-groups from the 2004 population. For the third and fourth iterations, only the numbers of enrolled students are used as an optimisation objective.

C.3.b Mathematical formulation

The numbers of enrolled students are calculated by converting the numbers of school-age population into numbers of enrolled students. The conversion process of school age-population into enrolled students require the following steps:

1. Defining typical age bands of elementary and secondary education in Saudi. These age bands for elementary and secondary are between (6-12) and (15-18) years old respectively.
2. Disaggregating the population counts that are based on five-year age-groups into counts of singular-ages of population. This disaggregation assumes a uniform distribution of ages within each five-year based age-groups. For example, if age-group (0-4) consists of 1000 person, then each age within this age-group (i.e., age 3) is estimated to have 200 persons.
3. Aggregating the population counts of singular-age groups using the early stated typical age-groups for elementary and secondary schools which are between (6-12) and (15-18) years old respectively.
4. Multiplying the aggregated population counts of school-age population by the ratio of enrolment level to get the final counts of enrolled students by education stage and gender.

The mathematical expression for converting the aggregated counts of school age population into counts of enrolled students can be written as follows:

$$\overline{Enrolled_x} = P'_x * E_x \tag{10}$$

where:

$\overline{Enrolled_x}$ The estimated total number of enrolled students in education stage x.

P'_x The estimated count of school-age population for education stage x.

E_x The estimated enrolment ratio of school-age population in education stage x.

To extend the expression further, the P'_x represents comprehensively the steps (1 to 3) mentioned earlier. Therefore, it involves listing all relative population age-groups for the stage x in a list called P_x. Then, each age-group (Age_i^x) is disaggregated by the size of the age group $C^{Age_i^x}$ and the number of years wanted from the age-group $W^{Age_1^x}$. Consequently P_x' is an aggregation of all relative schooling ages related to stage x listed in P_x. These steps are expressed as follows:

$$P_x = [Age_i^x, \dots, Age_n^x] \tag{11}$$

$$P_x' = \left(\frac{Age_1^x}{C^{Age_1^x}} * W^{Age_1^x}\right) + \left(\frac{Age_2^x}{C^{Age_2^x}} * W^{Age_2^x}\right) \dots + \left(\frac{Age_n^x}{C^{Age_n^x}} * W^{Age_n^x}\right) \tag{12}$$

$$P_x' = \sum\nolimits_i \left(\frac{Age_i^x}{C^{Age_i^x}} * W^{Age_i^x}\right) \tag{13}$$

where:

P_x A list of all age-group counts that is included in education stage x.

Age_i^x Counts of population in age-group i that are related to education stage x.

$W^{Age_i^x}$ The count of related ages within age-group i for education stage x.

$C^{Age_i^x}$ A constant equivalent to the size of age-group i for education stage x.

Finally, the extended mathematical expression of estimated enrolled students by education stage can be in this form:

$$\overline{Enrolled_x} = \sum\nolimits_i \left(\frac{Age_i^x}{C^{Age_i^x}} * W^{Age_i^x}\right) * E_x \tag{14}$$

After defining the estimation methodology for the number of enrolled students, the calibration process can evaluate the actual values against the model estimated values. The optimisation objectives were evaluated as follows.

The main calibration objective was set to minimise the difference between the estimated values and the actual values set to zero. This calibration function for the data curation model can be written as follows:

$$Min \sum\nolimits_x (\overline{Enrolled}_x - Enrolled_x) + (\overline{Age}_i - Age_i) \tag{15}$$

where:

$\overline{Enrolled}_x$ The estimated total number of enrolled students in education stage x.

$Enrolled_x$ The actual total number of enrolled students in education stage x.

$\overline{Age}_i$ The estimated count of population at age-group i.

Age_i The actual count of population at age-group i.

The calibration objectives included the numbers of enrolled students in elementary and secondary education in addition to selected population age-groups from the 2004 census. However, different combinations of objectives were tested such as the inclusion and exclusion of all 2004 census data to test for better results. Moreover, the calibration of the 1999 base year population was set as discrete values with 100 person step size. This is believed to enhance the speed of the process with minimum influence on the prediction's quality. The remaining parameters were considered as continuous values. Also, the number of iterations was set between 100,000 and 2.5 million iterations. The calibration time increased exponentially with the number of iterations, reaching more than 24 hours for 2.5 million iterations. Also, it was noted that after 300,000 iterations, the improvement in calibration goodness becomes marginal.

C.3.c The data curation results

To evaluate the calibration goodness, we used the measures Mean Average Percentage Error (MAPE), Standard Deviation (±SD) and Mean Absolute Deviation (MAD) for their ease of use and interpretation. It is possible to argue that, despite some notable and case-specific patterns of high margins of error, the curation method achieved reasonably low error margins. To describe the results of the curation model, first the number of enrolled students will be specified. After that, the selected age groups from the 2004 census are evaluated.

- The assessment of estimated numbers of enrolled students:

The overall level of error for both genders in all education stages over the years 1999, 2004, 2009 and 2014 averaged (MAPE, 3.80% ± 5.58%) and (MAD, 21681). Although, a below 10% MAPE margin of error could be accepted as a decent calibration result for 96 parameters, the examination of each education stage separately by gender and year revealed a pattern of increasing error in recent years for the secondary education stages. See Table 22 for the error levels produced by the curation model for the numbers of enrolled students disaggregated by stage, gender and year. Also, see Table 23 for the summary of error levels produced by the curation model for the numbers of enrolled students by stage and gender.

Table 22 The error levels produced by the curation model for the numbers of enrolled students disaggregated by stage, gender and year.

Education stage	Gender	Year	Actual values	Estimated values	Estimated enrolment ratios	Value of absolute error	Value of absolute percentage error
Elementary	male	1999	1188503	1189073	0.931	570	0.05%
		2004	1245237	1258308	0.955	13071	1.05%
		2009	1272086	1271648	0.98	438	0.03%
		2014	1318663	1394344	0.98	75681	5.74%
	female	1999	1096040	1095983	0.921	57	0.01%
		2004	1131617	1140639	0.932	9022	0.80%
		2009	1206334	1206184	0.98	150	0.01%
		2014	1301839	1312416	0.98	10577	0.81%
Secondary	male	1999	387720	387710	0.722	10	0.00%
		2004	474367	474456	0.746	89	0.02%
		2009	539924	489258	0.77	50666	9.38%
		2014	581057	528328	0.794	52729	9.07%
	female	1999	363668	360060	0.667	3608	0.99%
		2004	446578	446460	0.7	118	0.03%
		2009	506363	416804	0.732	89559	17.69%
		2014	565450	480403	0.764	85047	15.04%

The underlined values highlight the data with a high error margin

Table 23 The summary of error levels produced by the curation model for the numbers of enrolled students by stage and gender.

	Elementary male	Elementary female	Secondary male	Secondary female
MAD	22440	4952	25874	44583
MAPE	1.72%	0.41%	4.62%	8.44%
MAPE Std.	2.36%	0.40%	4.61%	7.99%

Table 22 shows that most numbers of enrolled students in elementary education had an absolute error percentage close to 0%. Only, for the 2014 male elementary stage, the absolute error percentage stood at 5.74% which is considerably higher than all other elementary estimates. Consequentially, Table 23 shows that the error levels for male students at their elementary stage jumped about 4 folds to (MAPE, 1.72% ± 2.36%) (MAD, 22440) when compared to the level of elementary female education which is standing at (MAPE, 0.41% ± 0.40%) (MAD, 4952). In this case, it could be argued that male fertility rates might be slightly higher than what they should be if the numbers of actual enrolled students were held to be accurate. Moreover, it is worth mentioning that calibrated enrolment ratios were mostly biased towards the higher calibration bands. Therefore, it is possible that a lower enrolment ratio affected the more recent period between 2009 and 2014.

On the other side, the error level of enrolled students in secondary stage increased noticeably from the year 2009. For the years 1999 and 2004, the absolute percentage error was close to 0%. However, for the years 2009 and 2014, both male and female absolute percentage error were significantly higher, averaging 9.32% and 16.36% respectively. Accordingly, the overall secondary stages level of errors for male and female were (MAPE, 4.62% ± 4.61%) (MAD, 25874) and (MAPE, 8.44% ± 7.99%) (MAD, 44583) respectively. Both MAPE and MAD highlights a substantial increase in the error levels caused by the underestimation of enrolled students for recent years. Moreover, the secondary stage has a higher level of error when compared to the error level of elementary education. Furthermore, when comparing the error levels of male and female enrolled students of the secondary stage, it is noticeable that the female level of error is more than two fold when compared to their male counterparts. These two aspects might be due to: 1) a higher levels of enrolment ratios than expected for the secondary stages; 2) a narrower gap in gender education than anticipated. Regarding these assumptions, all enrolment ratios outcomes were biased towards the maximum calibration bands. The Kingdom of Saudi Arabia has witnessed a massive increase in the provision of educational facilities in the last two decades. For example, the number of secondary schools in Saudi Arabia jumped from 2938 schools in year 2000 to 4909 schools by 2010 (UNESCO-IBE, 2011), which accounts for about 70% of the increase in the number of secondary schools in 10 years. Moreover, these schools were almost equally divided between both genders. The author believes that the reported net enrolment ratios for secondary education does not reflect the increase in the number of schools. While the number of schools increased by 70% between year 2000 to 2010, the number of enrolled students increased by 40% for the same period. Although the number of schools is not expected to correlate perfectly with their total capacity, the size of schools' provision would facilitate higher enrolment ratios at the kingdom level. Moreover, unlike the elementary stage which is calculated using age-groups (5-9) and (10-14), the secondary stage is calculated based on age-group (15-19) only. Therefore, the counts of elementary stages are based on two rounds of birth counts. This means that elementary counts are more constrained, yet they have a lower margin of error for all years than the secondary stages, which has an underestimation issue only in recent years. All these reasons point towards a possible higher and equal enrolment ratio for both genders in the secondary education stages. If the curation model predicted the population counts of age group (15-19) with accuracy and the average numbers of enrolled students has also a high accuracy level, then the expected male and female net enrolment ratios for year 2009 should be around 85% and 89% respectively. Also, for the year 2014 the expected net enrolment ratios are 87% and 90% respectively.

- The assessment of population counts:

For the calibration of 1999 dataset, the error levels for male and female were (MAPE, 13.86% ± 8.67%) (MAD, 53320) and (MAPE, 16.44% ± 11.07%) (MAD, 68711) respectively. It is worth noting that, the calibration of female population in childbearing years was more biased towards the minimum limits. This might indicate that the higher levels of AS-CFR comparing to AS-TFR resulted in a decrease in the number of females. However, despite the decrease of women numbers in childbearing years in year 1999 by 471870 women, we notice that that calibrated women numbers in childbearing years in year 2004 were higher than the official reported numbers by 378864 women. See Table 24 for error levels of estimated 1999 population counts by age group and gender.

Table 24 error levels of estimated 1999 population counts by age group and gender.

Age-group	Error value (male)	Percentage of Absolute error value (male)	Error value (female)	Percentage of Absolute error value (female)
0--4	-203571	18%	-219673	20%
5--9	-98561	9%	-168697	18%
10--14	74933	7%	104123	10%
15--19	86737	10%	82946	9%
20—24	130506	18%	-171218	33%
25--29	50341	9%	-138013	33%
30--34	63495	13%	-110044	33%
35--39	48194	12%	-89315	32%
40--44	27509	9%	-10169	4%
45--49	8974	4%	-18339	9%
50--54	-16497	11%	-17718	11%
55--59	3091	2%	-2687	2%
60--64	8754	8%	-5194	5%
65--69	-31879	33%	-294	0%
70--74	-21907	33%	-9053	20%
75--79	13887	20%	8260	20%
80+	17610	20%	12345	20%

Moving into the estimation of year 2004 population, the selected age-groups from the 2004 census as direct calibration objectives belongs to, male population aged over 24 and, female population aged over 45. The direct calibration in this context means that, the beforementioned age-groups were forced to match the 2004 census data, whereas the remaining age groups are calibrated indirectly using the rules and inputs of the CCM method and the number of enrolled students. For example, age-group (0-4) is calibrated implicitly using AS-FRs, counts of females in childbearing years between (15-55) and the number of enrolled students in the subsequent years. To evaluate the optimisation outcome, error levels for direct and non-direct objectives will be separated. See Table 25 for error levels of estimated 2004 population counts by age group and gender.

Table 25 error levels of estimated 2004 population counts by age group and gender.

Age-group	Error value (male)	Percentage of Absolute error value (male)	Error value (female)	Percentage of Absolute error value (female)
0--4	-22990	2.1%	20525	1.9%
5--9	11253	1.0%	58582	5.3%
10--14	19884	1.9%	204075	17.7%
15--19	-111293	10.5%	-124018	13.2%
20—24	-131354	14.7%	-109682	13.9%
25--29	-42	0.0%	188894	26.9%
30--34	30	0.0%	162722	28.3%
35--39	32	0.0%	166980	33.5%
40--44	-45	0.0%	94041	25.2%
45--49	-47	0.0%	-40	0.0%
50--54	36	0.0%	-33	0.0%
55--59	37	0.0%	21	0.0%
60--64	39	0.0%	-14	0.0%
65--69	-24	0.0%	-29	0.0%
70--74	-11650	13.1%	21	0.0%
75--79	-4168	7.8%	29	0.1%
80+	5946	9.9%	20151	36.1%

Directly calibrated values

Table 25 shows that, on one hand, for male population, the directly calibrated age-groups had almost 0% error percentage for the age-groups between age 25 and 69. However, for the remaining age-groups which are between age 70 and above 80, the error levels were higher at (MAPE, 10.2% ± 2.2%) (MAD, 7254). The higher level of error in the elder age-groups does not influence the targeted school-age population and it can be sorted by adjusting the calibration limits of 1999 dataset. For the non-directly optimised age-groups, the error levels were at (MAPE, 6.0% ± 5.5%) (MAD, 59355) while the error level of age-groups (15-19) and (20-24) had (MAPE, 12.6% ± 2.1%) (MAD, 121324). The error level of age-groups under age 15 were at (MAPE, 1.7% ± 0.5%) (MAD, 18042). On the other hand, for female population, the directly calibrated age-groups had almost 0% error percentage for the age-groups between age 45 and 79. Only age-group Above 80 had an absolute error percentage of 36.1%. Like the male population, the high percentage of error is irrelevant to the targeted school-age population. The error level for the non-direct optimisation of age-groups between age 0 and 44 was at (MAPE, 18.4% ± 10.2%) (MAD, 125502). Therefore, it can be said the direct optimisation was successful to a great extent in achieving its objectives. Moreover, the higher levels of error in the non-directly optimised age-groups were expected to correct the data and achieve low error levels in the numbers of enrolled students.

- The assessment of fertility rates

Given the strong constraints that were imposed on the calibration of fertility rates, the differences between the actual and calibrated rates were marginal. The calibrated CFR averaged (MAPE, 0.68% ± 0.72%) (MAD, 0.037) for both genders over the years 1999, 2004 and 2009. See Table 26 for error levels of estimated CFR by year and gender.

Table 26 error levels of estimated CFR by year and gender.

year	new-born gender	Actual	Calibrated	Value of absolute percentage error
1999	Male	2.877	2.942	2.23%
2004	Male	2.853	2.853	0.01%
2009	Male	2.689	2.696	0.26%
1999	Female	2.709	2.738	1.07%
2004	Female	2.680	2.680	0.01%
2009	Female	2.565	2.577	0.46%
1999	Both	5.586	5.680	1.67%
2004	Both	5.533	5.533	0.00%
2009	Both	5.254	5.273	0.36%

Table 26 shows that calibrated CFR values are slightly higher than actual values. Moreover, the monotonic decreasing trend is present in both actual and calibrated values. However, the monotonic decrease in calibrated values seems to be a bit slower than actual values.

On the other hand, the calibrated AS-FRs mirrored the shifting trends in decreasing fertility and moving peak in reproduction to older ages. See Figure 20 for the calibrated AS-CFR for both genders births for the period between 1999 and 2014.

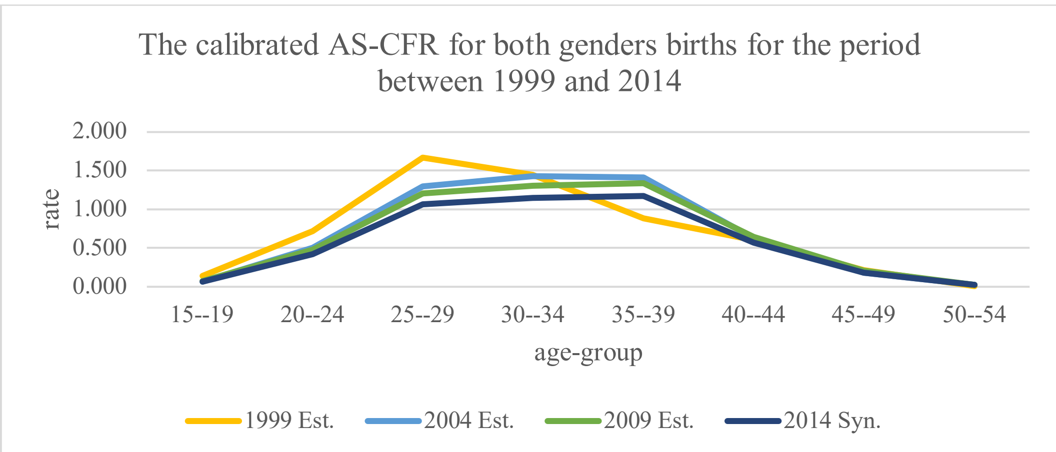
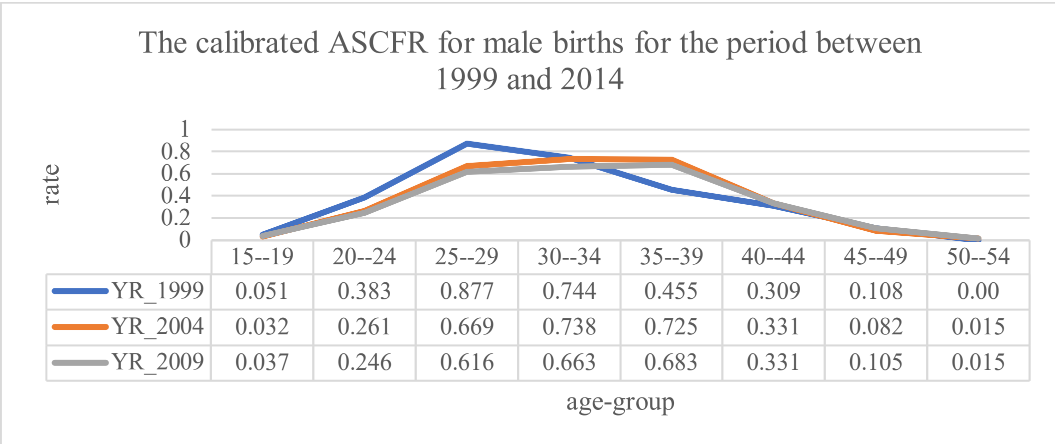


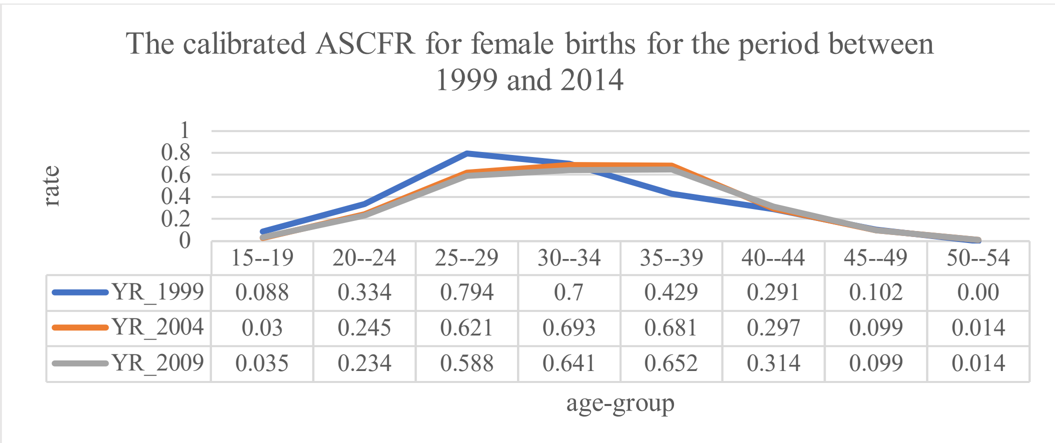
Figure 20 the calibrated AS-CFR for both genders births for the period between 1999 and 2014.

Given that there are no actual values for AS-CFR, the assessment of error levels cannot be done by comparing actual and estimated values, instead we can compare the synthesised AS-FR against the calibration AS-FR, which in this case represents comparing estimations against each other. The author believes that such a comparison will not add any instructive information to the discussion. Therefore, comparing the actual and estimated CFR is sufficient at this stage. However, in terms of the fertility distribution on ages, better insights could be drawn with comparing the calibrated AS-CFR in Figure 20 with its calibration limits in Figure 17. It can be noted from comparing the AS-CFR with its calibration limits, that the calibration outcomes did not increase the rate for age group (25-29) as the calibration limits allow or suggest. This outcome might indicate that the AS-TFR was sensitive to some changes in years 2004 and 2007 for age-group (25-29) which was corrected by the model. Moreover, both calibrated rates are very similar in terms of shifting trends and distribution. For disaggregated values of calibrated AS-FRs, see Figure 21 for the calibrated AS-CFR for male births for the period between 1999 and 2014 and Figure 22 for the calibrated AS-CFR for female births for the period between 1999 and 2014.



	15--19	20--24	25--29	30--34	35--39	40--44	45--49	50--54
YR_1999	0.051	0.383	0.877	0.744	0.455	0.309	0.108	0.00
YR_2004	0.032	0.261	0.669	0.738	0.725	0.331	0.082	0.015
YR_2009	0.037	0.246	0.616	0.663	0.683	0.331	0.105	0.015

Figure 21 calibrated AS-CFR for male births for the period between 1999 and 2014.



	15--19	20--24	25--29	30--34	35--39	40--44	45--49	50--54
YR_1999	0.088	0.334	0.794	0.7	0.429	0.291	0.102	0.00
YR_2004	0.03	0.245	0.621	0.693	0.681	0.297	0.099	0.014
YR_2009	0.035	0.234	0.588	0.641	0.652	0.314	0.099	0.014

Figure 22 calibrated AS-CFR for female births for the period between 1999 and 2014.

C.3.d Remarks on the model and its outputs

The curation model represents a serious step towards fixing the data issues found at the Kingdom level of Saudi Arabia. Moreover, the workflow of the model follows the natural change order of any population although constrained by the external references to enhance the accuracy of population numbers. Despite the obstacles imposed by lack of data, the model achieved promising results in correcting the population numbers using data references with credibility and within defined calibration limits. The improvements to the model outcomes could be done by adding more reference data and enhanced calibration limits. This model could thus be applied to any smaller demographic level. However, this is conditioned by the availability of data. For the level of Riyadh city, the data is critically lacking and therefore, the model was not tested, and the curation results were simply applied to the city level.

D. Future forecast

For the purpose of producing fertility projections, five extrapolation methods were tested. These methods include LC (Lee and Carter, 1992), BMS (Booth, Maindonald and Smith, 2002), FDM (Hyndman and Shahid Ullah, 2007), ARIMA (random walk with a drift), and BATS (Livera, Hyndman and Snyder, 2011). While these models were applied using the R programming language, the first three models were applied using the Demography package (Hyndman *et al.*, 2019) and, the remaining two models were applied using the Forecast package (Hyndman *et al.*, 2020). Although the error measurements should be the main determinant for choosing the best forecasting method, the outcomes of the different error measurement can be viewed as misleading. This is not to say that the error measurements are not valid but to indicate that the demography package outcomes were unacceptable and that its inconsistency would not be identified if it was only relied on error indices. See Table 27 for averages of error levels across ages by forecasting model and birth’s gender.

Table 27 Averages of error levels across ages by forecasting model and birth’s gender.

Gender	method	ME	MSE	MPE	MAPE
Male	FDM	1.484	2.370	-1.730	1.730
	BMS	0.000	0.005	0.011	0.066
	LC	-0.022	0.091	0.021	0.160
	ARIMA(0,1,0)	0.000	NA	-2.346	9.478
	BATS	-0.004	NA	-2.453	9.300
Female	FDM	1.471	2.338	-1.883	1.883
	BMS	0.000	0.008	0.021	0.086
	LC	-0.002	0.055	0.018	0.154
	ARIMA(0,1,0)	-0.002	NA	-2.362	9.309
	BATS	-0.005	NA	-3.093	9.408

Table 27 shows that the error measurements for both genders are almost identical. Moreover, the demography package models LC, BMS and FDM have a considerably low MPE and MAPE margins of error when compared to the ARIMA and BATS models. However, it was noticed that all the demography models had a higher forecasting jump-off and values for the lower forecasting band in some ages than their jump-off points for the upper forecasting bands. Therefore, the Demography package models were excluded. See Figure 23 for the initial jump off problem in projecting male birth rates using LC method.

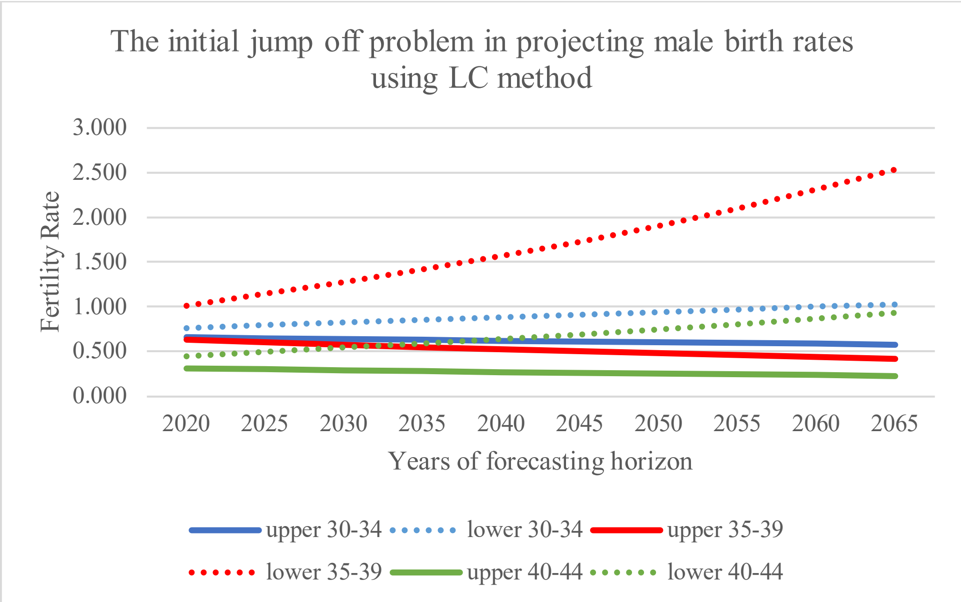


Figure 23 The initial jump off problem in projecting male birth rates using LC method.

Figure 23 shows an example for the issue of higher jump-off points and projection values of the lower forecasting bands than its upper forecasting bands. The main cause of this issue is unknown to the author, and it might be due to conflict in R packages and/or an unknown error in the package. Nevertheless, the causation of this issue is outside the scope of this research. Therefore, the choice of adopted fertility rates was between ARIMA and BATS models. The best alternative might be the BATS model because the BATS middle forecasting band shows some sort of stability around the historic average of complete fertility rates with a change in the structure of reproduction ages, whereas the rates projected by ARIMA model for the middle band were fixed through the years and do not reflect the possibility of structural changes, at least for the middle forecasting band.

Although the average rate for the BATS model increases monotonically from 4.86 to 5.31 births per women for the period between 2020 to 2065, the mean percentage of change from the average of historic data for the projected data and the historic data itself was 4% and 5% respectively. Moreover, despite that, an increase in the average rate might not be expected in general but the values are within its historic bands from 1975 to 2015. Therefore, this increase should be acceptable as a possible scenario although not likely to happen. See Figure 24 for the projection of complete fertility rate for the period between 2020 and 2065 using the BATS model.

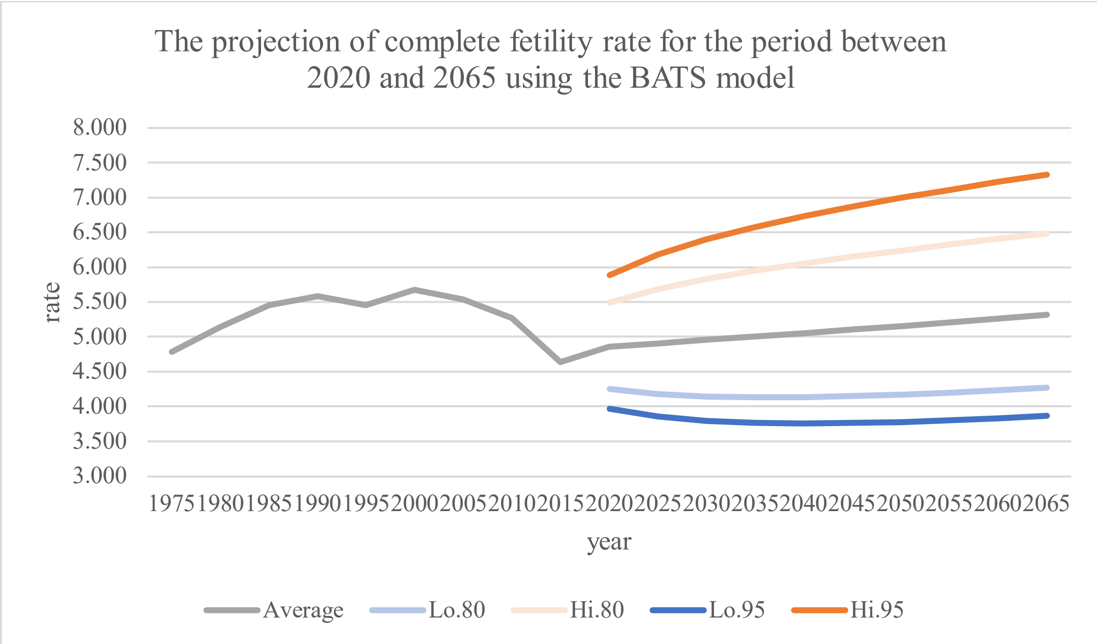


Figure 24 The projection of complete fertility rate for the period between 2020 and 2065 using the BATS model.

First, Figure 24 shows the BATS average points of forecast with two upper and lower bands covering the 80% and 95% confidence levels. Given that an increase in the average rate is unlikely, the author will use the lower band with 95% confidence level as the main input for the developed models and consider that remaining possibilities for further scenario testing. The accompanying change in the AS-FR can be found in Figure 25 which provides the average AS-FR projections for both birth's genders using BATS model for the period between 2020 -2065. Also, Figure 26 for the average AS-FR projections for both birth's genders using ARIMA model for the period between 2020 -2065.

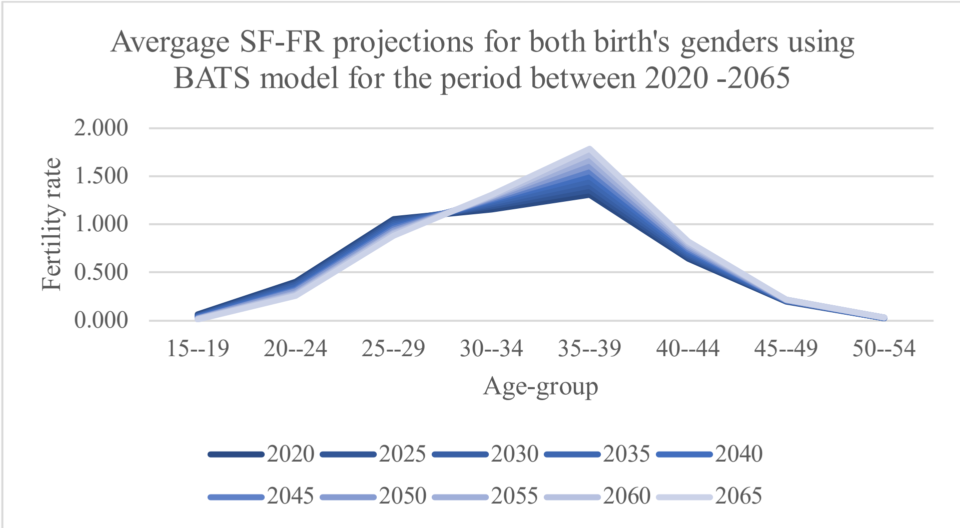


Figure 25 Average AS-FR projections for both birth's genders using BATS model for the period between 2020 -2065.

Figure 25 shows the continuous shift in fertility rates across all age groups. The main trend can be described as decreasing fertility rate for ages below 30 and increasing for ages above 30. Moreover, the peak reproduction will be age-group (35-40) with around two births in year 2065. In order to compare the average rates of BATS and ARIMA. See Figure 26 for the average AS-FR projections for both birth's genders using ARIMA model for the period between 2020 -2065.

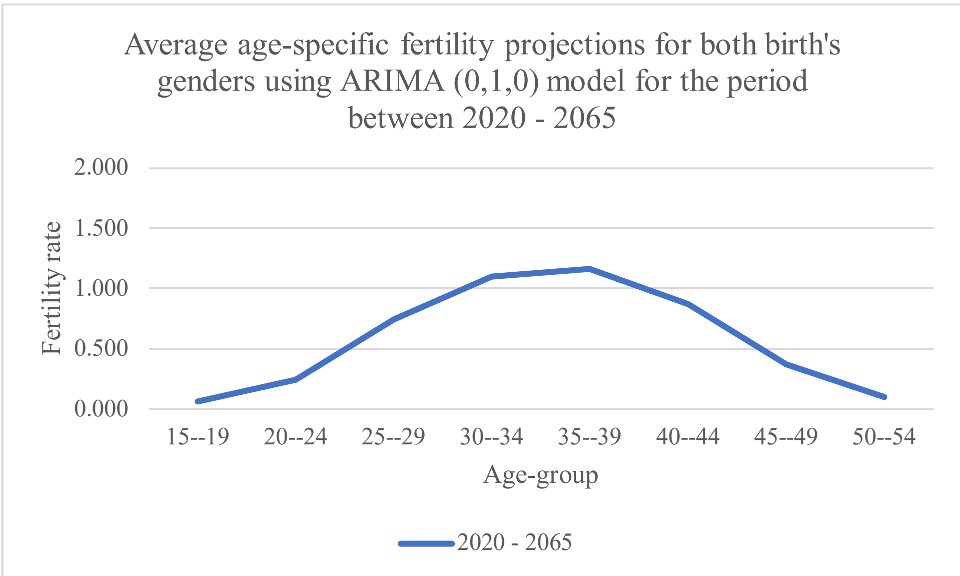


Figure 26 Average AS-FR projections for both birth's genders using ARIMA model for the period between 2020 -2065.

Figure 26 shows that the average fertility rate for the period between 2020 and 2065 is expected to remain static. Consequently, ARIMA was less preferable when compared to the BATS model.

Second, the BATS model has the advantage of not producing negative values. For both lower confidence bands, some age-groups in the ARIMA model reached below zero values from its jump-off year. For example, age-group (15-19) starts at -0.004 in 2020 and reaches -0.150 in 2065. This issue can be sorted by setting a lower limit but, in this case, the BATS model seems to be outperforming the ARIMA model. See Figure 27 for the lower band AS-FR projections for both birth's genders using BATS model for the period between 2020 – 2065. Also, see Figure 28 Lower band AS-FR projections for both birth's genders using ARIMA (0,1,0) model for the period between 2020 – 2065.

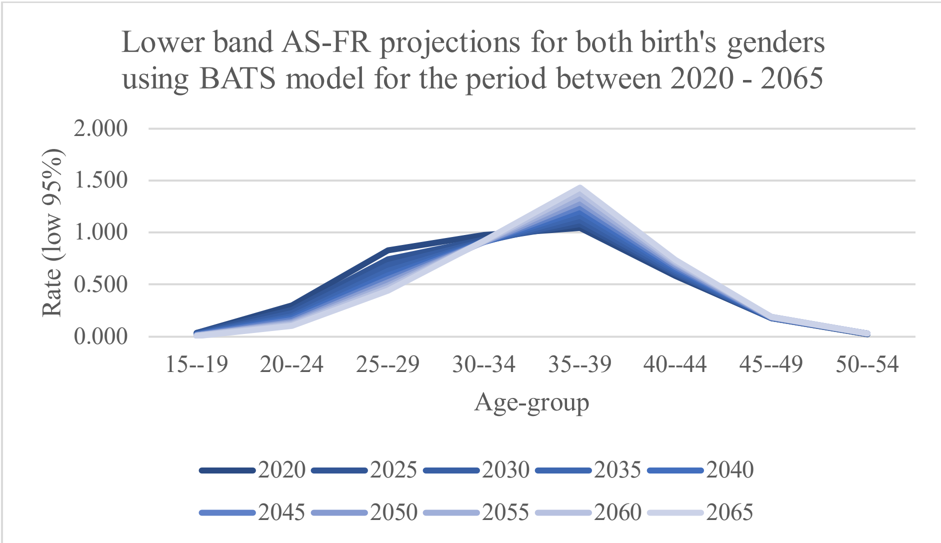


Figure 27 Lower band AS-FR projections for both birth's genders using BATS model for the period between 2020 – 2065.

Figure 27 Shows that the BATS rates are always above zero. Moreover, the shift in peak reproduction is present at the lower forecasting band.

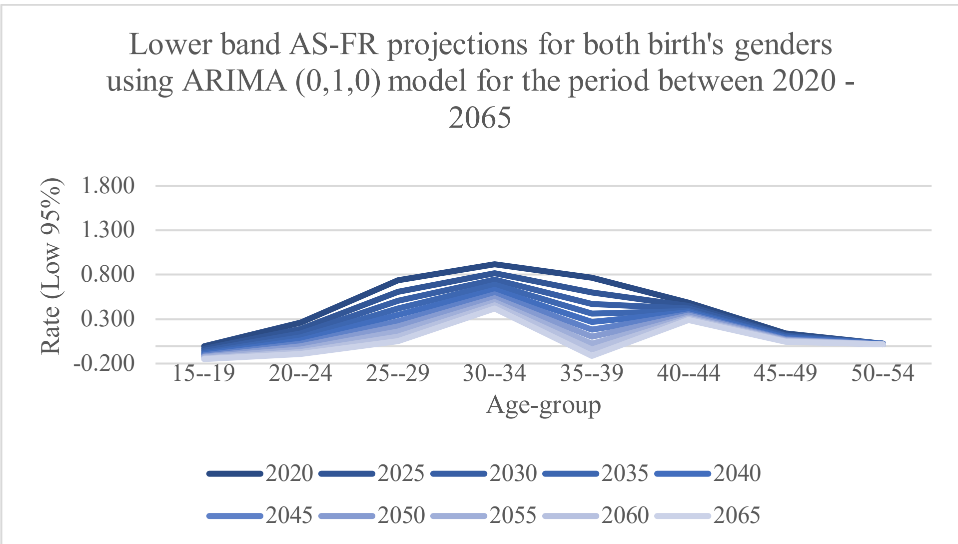


Figure 28 Lower band AS-FR projections for both birth's genders using ARIMA (0,1,0) model for the period between 2020 – 2065.

Moving into the ARIMA lower band, we can notice from Figure 28 that some age-groups reach sub-zero values in a rapid manner, Moreover, the distribution of births is clearly unrealistic as we move towards 2065, considering that age-group (35-39) will decrease to below zero, while age-group (40-44) will have a birth rate that is close to the peak reproduction in age-group (30-34). Therefore, it is believed that the ARIMA model in its current specification is not suitable for consideration and requires further refinement.

Given the aforementioned reasons which include having an average projection that reflects the continuity of historic values, the trend of shifting birth distribution and the constant above zero rates is extended. The BATS model is closer to be the ideal option than the other five models. More specifically the lower band with 95% confidence level which inherited its features from the fertility data inputs. See Table 28 for the lower band BATS model outcomes with 95% confidence level by age-group and year.

Table 28 The lower band BATS model outcomes with 95% confidence level by age-group and year

Age-group / year	2020	2025	2030	2035	2040	2045	2050	2055	2060	2065
15--19	0.036	0.031	0.027	0.023	0.020	0.017	0.014	0.012	0.010	0.009
20--24	0.301	0.256	0.223	0.197	0.175	0.157	0.141	0.127	0.115	0.104
25--29	0.826	0.745	0.684	0.634	0.591	0.553	0.520	0.491	0.464	0.440
30--34	0.979	0.935	0.917	0.909	0.907	0.907	0.909	0.912	0.915	0.918
35--39	1.047	1.085	1.124	1.164	1.206	1.249	1.293	1.339	1.386	1.435
40--44	0.582	0.599	0.616	0.633	0.650	0.668	0.685	0.703	0.720	0.738
45--49	0.173	0.175	0.177	0.179	0.181	0.183	0.185	0.186	0.188	0.189
50--54	0.026	0.027	0.027	0.027	0.028	0.028	0.028	0.028	0.029	0.029
Total	3.970	3.853	3.795	3.766	3.757	3.762	3.776	3.798	3.827	3.862

3.8.1.3 Population based on household data (at the districts level)

The only source of household data at the district level for this study was the ADA 2016 demographic survey, even though a more recent survey was carried out around the year 2020. Furthermore, the sample size was predetermined to cover 1.5% of all Riyadh city residential units which result in a total number of 16508 surveyed households. Furthermore, both response rate and confidence level were acceptable at 97.5% and 95% respectively. Moreover, the survey was designed to cover multiple aspects such as social, education, residential mobility, economic and migration rates. This survey is believed to provide accurate and full information about the current situation in Riyadh city. Therefore, the outcomes of these surveys are considered as a foundation for plan drawings, policies forming, and decisions made by ADA. To assess the data quality and its suitability for the purpose of this research, the following steps were undertaken:

A. Description of original data format at the district level

A.1 Population counts

Population data within its household composition will be used for the districts level model. This data is more complex than the individual's data that was described in the city level data. While this data provides demographic characteristics about the household head that includes age, gender, education level, employment status, marital status and the number of family members, the data provides less information about the household members such relationship to household head, age and gender. See Table 29 for the demographic characteristics for the population data based on households.

Table 29 the demographic characteristics for the population data based on households.

Data source	Data description	range of sub-classifications		Number of sub-categories
Demographic data	The number of children in each household	**From 0 up to 20 child**		21
	The age of each household member	**0-4**	**5-9**	17
		10-14	**15-19**	
		20-24	**25-29**	
		30-34	**35-39**	
		40-44	**45-49**	
		50-54	**55-59**	
		60-64	**65-69**	
		70-74	**75-79**	
		80+		
	The gender of each household member	**male**		2
		female		
	Educational level of household head	**Literacy**		8
		Reading only		
		Reading and write		
		Primary		
		Intermediate		
		Secondary		
		University		
		Postgraduate		
	Marital status of household head	**divorced**		4
		married		
		single		
		widowed		
	Work status of household head	**employed**		7
		Home wife		
		retired		
		student		
		Unable to work		
		Unemployed not searching		
		Unemployed searching		

Table 29 shows the demographic characteristics available for the households in Riyadh city. It is worth mentioning that the age-group classification is similar to GAS data, and also the data has zero missing responses when it comes to demographic information.

With regard to the population change rates, which include the rates of fertility, mortality, and social change (such as marriage and divorce), the data is lacking at the district level. As a result, and in accordance with what was employed in the data pertaining to the city level, synthesised rates from the Kingdom level will be utilised at the district level. The fertility and mortality rates that will be used are those that have been calibrated based on the data from the city level. The Marriage and Divorce Statistics (GAS, 2020) are the only data source that is currently accessible, and it only gives general rates that are not disaggregated by age group or any other demographic factors. The rates of marriage and divorce for Saudi population of adults aged 15 and older are, respectively, 9.6 and 3.64 per 1000 people.

On the other hand, the ADA 2016 survey is intended to include topics relating to migration as well as residential mobility, and it does so in the context of zonal growth rates. Despite the fact that migration data were not supplied, the data that was used was propagated from the city level crude migration rate. Within the framework of the demographic survey, there were a total of six questions devoted to the topic of residence mobility. These questions include the respondent's current district, their prior district, the number of years they have lived in their present residence, the annual rent they pay, and the reason they moved most recently. See Table 30 for the data available for residential mobility in Riyadh city with its categories of answers and number of missing responses

Table 30 the data available for residential mobility in Riyadh city its categories of answers and number of missing responses.

Data source	Data description	Range of sub-classifications	Number of sub-categories	Missing responses
Residential mobility data	Current district	from 0 to 150	150	0
	Previous district	from 0 to 150	150	4185
	Residence ownership	**Other**	3	0
		Owned		
		Rented		
	Annual rent	From 1200 up to 250000 SAR	2488	0

Table 30 the data available for residential mobility in Riyadh city its categories of answers and number of missing responses (continued)

Data source	Data description	Range of sub-classifications	Number of sub-categories	Missing responses
Residential mobility data	Number of years lived in current residence	From 0 up to 75 years.	76	0
	Reason for movement	**Bought a home**	15	3488
		Building demolish		
		Distance to transportation		
		Distance to family		
		Distance to markets		
		Distance to school		
		Distance to work		
		District satisfaction		
		High rent		
		Family size		
		Improved income		
		Lack of services		
		Other		
		Owner reclaimed		

Table 30 shows that the response rate for some questions are severely affected and this issue would reflect on the data quality as will be discussed in the coming sub-section.

A.2 Data quality evaluation

In terms of contextual data quality, according to GAS (2016) , five demographic surveys have been carried for Riyadh city over the past 30 years, with the most recent survey in 2016. Moreover, these demographic surveys were designed to be comprehensive and cover multiple aspects such as social, education, residential mobility, economic and migration rates. Also, it is believed to provide accurate and full information about the current situation for the future development of Riyadh city. Therefore, the outcomes of these surveys are considered as the foundation for policy forming and decision making by ADA. However, it is worth noting that all prior surveys to 2016 focused on the municipal level and above. Only the 2016 survey aimed to focus on the differences at the district level.

Regarding the sampling process of the 2016 survey, only basic information was provided about the sample design which states that stratified random sampling was applied. Moreover, the sample size was predetermined to cover 1.5% of all Riyadh city residential units which resus in a total number of 16508 surveyed households. In addition, the survey intended to target both Saudis and non-Saudis. Unfortunately, further information about the expected representation level from the sample design for the different investigated aspects was not provided. For example, it is unknown if the sample was intended to distinguish between the trends of households migrating to Riyadh city by its size and income or just aiming for the crude number of migrants.

Concerning the compatibility of the data, the information was disaggregated in the same manner as the data at the city level. Therefore, there were no problems encountered with linking the data. Furthermore, the presented data was recent, while in the case of residential mobility, the data included mobility movements dating back 75 years. In addition, with regard to the regularity of the data collection, the data was not collected in a consistent manner, and it was only collected five times during the course of the previous 30 years. Aside from that, the author was only able to obtain the most recent survey. Therefore, in terms of the consistency of the data in the classifications and definitions, there was no similar data available in order to establish a comparison with the data's consistency.

When it comes to Intrinsic DQ, given that the data is coming from an official body, it is conceived with a high level of confidence. In addition, other than the broad information regarding the response rate and the level of confidence, there was no detailed error reporting that was provided. For the purpose of the DQ concerning representational and accessibility, the data is easily readable and were delivered in a digital format that is reasonable to edit and adapt. Furthermore, publishing policies are stringent, yet the data was exclusively provided for the purpose of this research without permission to disclose or share it in its raw form.

B. Data preparation at the district level

B.1 Data Exploration at the district level

Before the demographic data can be used as an input for the model at the district level, there are four problems that need to be fixed that were discovered in the data. These concerns are related to the fact that some data were missing, the amount to which mobility trends were captured, the coverage level of all potential household types, and the weights of sampling which are not evenly distributed. The following is an explanation of each of these aspects:

B.1.a The issue of missing data

Although, all 9160 respondents completed the demographic part of the questionnaire, the residential mobility part was lacking in two aspects: the previous district location and the reason for movement. By removing all households which lacked responses, the sample size plunges by 45.86% yielding only 4959 households with complete responses. Moreover, despite the hope in finding valuable insights from the prolonged coverage period in residential mobility data, this is adversely affected by the distribution of sample size over the years and the number of households with complete response. By and large, the sample size of each year decreases when moving towards 1941. To illustrate the sample size of older residential mobilities, there are only 80 complete surveys for the period between 1941 and 1980. This is another more recent example for the sample being affected comes from the year 2016 which had 2777 households but only 168 with complete responses. See Figure 29 for the sample size of residential mobility in each year for the period between 1941 to 2016.

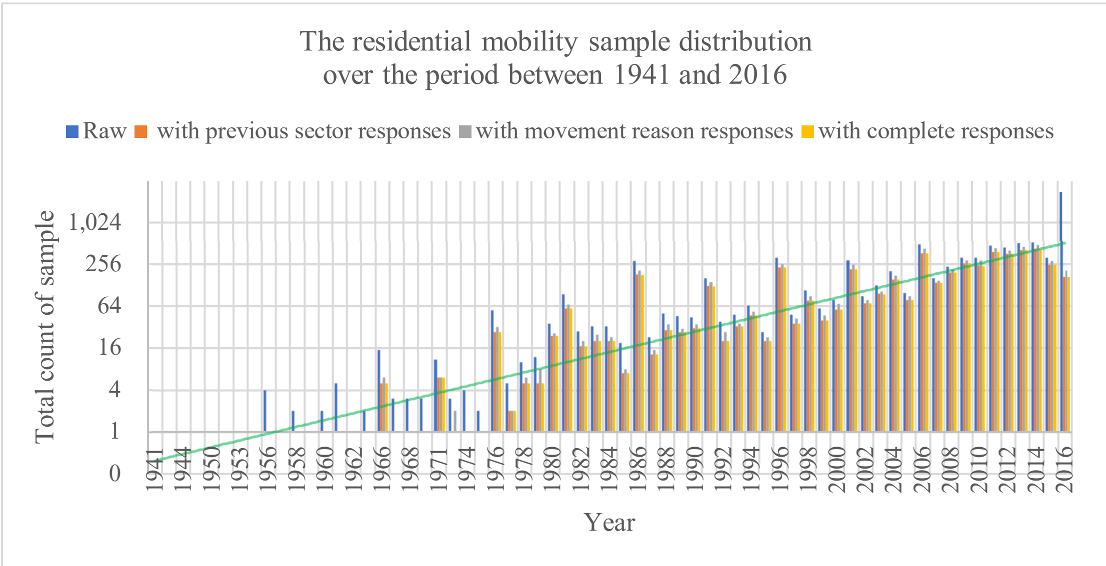


Figure 29 the sample size of residential mobility in each year for the period between 1941 to 2016.

As can be seen from Figure 29, the year 2016 is the most affected year by missing responses in comparison to the past 10 years at least. While the average percentage of missing responses for the past 10 years is around 20% with a standard deviation of 3%., in year 2016 alone the missing responses affected about 94% of the surveys. Also, it is worth mentioning that the number of moving residents in 2016 consists of about 30% of the total sample size. This aspect raises some concerns over the data accuracy especially for year 2016. Unfortunately, the data does not distinguish between the moves as if they originated from residential mobility or newly migrated residents to the city and this narrows our assumptions about the cause of this inconsistency in the data. Nevertheless, on the one hand, considering that about one third of Riyadh city population had changed their residence location in 2016 is unreasonable. On the other hand, considering that any missing respondents belong to newly migrated residents is also irrational as it would mean that Riyadh city population might have increased by over one million residents in a single year. Moreover, if we take the average sample size of the past 10 years, we can find that the percentage of residential mobility is averaging about 5% with a standard deviation of 2%. These rates are more consistent and reasonable to be taken into account than those of year 2016. Therefore, it would be more sensible to consider filtering and aggregating the samples of year 2016 in the data.

B.1.b The coverage level of all possible household types.

In an ideal sampling practice, all targeted populations are covered in some way. However, given the number of factors available in the household's survey and its disaggregated subcategories that were mentioned earlier in Table 29, the number of possible unique household types could reach 159,936 types. Therefore, it becomes clear that not all types of households would be covered by this sample, which consist of randomly selected 9160 households. Moreover, to illustrate this problem with an example from the perspective of residential mobility, in the evaluation of residential moves that occurred between 2005 and 2015, four demographic aspects were considered. Three aspects related to the households' head, his or her age, work status and education level. In addition to the households' size, it was found that the mean number of available unique household types moved between any two districts was 1.9 household types with a standard deviation of 2.48. Moreover, the maximum number of unique types moved among any two districts was 40 household types. Consequently, the coverage of all households by type is far from Ideal and in most cases, there might not be enough samples to the same type to train a prediction model. This issue draws the author's attention to investigate the availability of mobility trends in the data that is suitable to establish a prediction model.

B.1.c The extent of capturing mobility trends

One of the main features that must exist in the data is non-random mobility behaviour, in other words, the availability of mobility trends that can be replicated. For the purpose of examining the existence of this feature among the district's boundaries of Riyadh city, a circular migration flow plot was used to visualize residential mobility trends. Abel and Sander (2014) suggest that this method provides a legible quantitative visualisation of entire bilateral flow tables. Therefore, a residential mobility flow plot was created to illustrate the mobility patterns. The plot was generated using Circlize package in R (Gu *et al.*, 2014). In this plot each zone is using a unique colour. In addition, the arrows illustrate the flow direction. Also, the width of the lines represents the flow size. Moreover, the sections' axis shows the name of the municipality with the migration size in numbers. The value of each tick mark on the axis is equivalent to 10 households. The flows illustrated in this plot are based on the period between 2010 and 2015. In order to simplify the complexity of flows to increase the plots legibility, these flows are aggregated at the municipal level and consider the count of households without their unique characteristics. See Figure 30 for the residential mobility pattern among Riyadh city districts for the period between 2010 and 2015 using circular flow diagram.

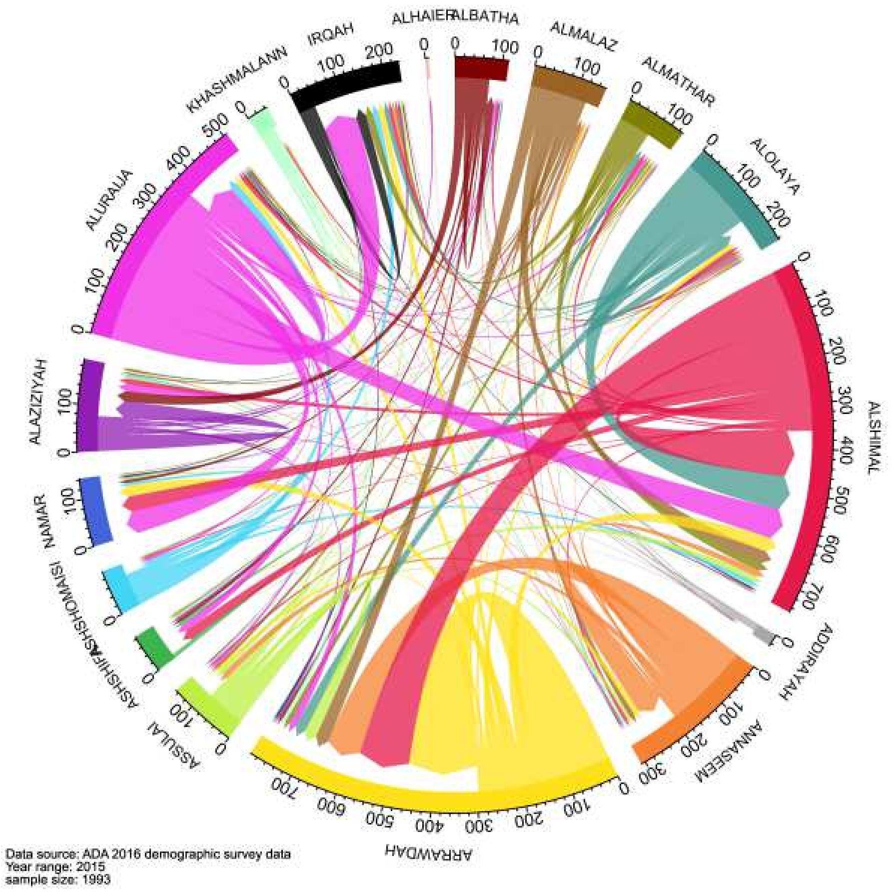


Figure 30 the residential mobility pattern among Riyadh city districts for the period between 2010 and 2015 using circular flow diagram.

As can be seen from Figure 30, there is an obvious residential mobility pattern among the municipalities of Riyadh city. For example, Alshamal, Arrawdah , Irqah and Namar municipalities can be viewed as main attractors of residential mobilities across the city. Moreover, both Alshamal and Arrawdah have a large internal mobility movement. On contrary, Aluraija, Almalaz, Ashshomaisi and Alolaya have a repulsive effect on their residents, although these municipalities still attract some mobilities. One particularly interesting observation is that Namar municipality has no out-flow mobility.

The availability of mobility trends was tested for every past five years and showed the existence of mobility trends in the data at least at the highest level of demographical and geographic aggregation.

B.1.d The nonequal weights of sampling.

Due to the nature of sampling in capturing the distribution/occurrence of investigated events or phenomena, the collected sample becomes associated with weights that reflect the actual occurrence of that event or phenomena. For example, if we have 130 researchers in the urban system and 40 of whom are working in urban simulation then, a sample of 10% would mean 13 participants. If the sampling process was distributed in a perfect manner, then we would get 4 participants specialised in urban simulation out of the 13 participants. Note that in this case the 10% ratio remains constant. Therefore, each respondent represents 10 researchers in real life including those working in urban simulation. However, if the sample distribution was not perfect then, for example, we could get 1 researcher specialised in urban simulation. This would mean that the 10% conversion rate is no longer applicable and requires some sort of adjustment. In this case the 1 researcher should be weighted as 40 and each researcher of the remaining 12 respondents will be equal to 7.5 researchers in real life to account for the 130 targeted population. For the purpose of this research, the sample data will be used in an ABM as a base population. Consequently, the ABM agents will be created by the sample population, which then will interact with each other in forming new households through events such marriage, divorce and births. Unlike the common practice of synthesising sample population data into full population, this research will use the sample data primarily, then convert the outcomes into population counts using the original conversion weights of the sample. The issue here is that the predetermined 1.5% sampling target is not expected to result in a fixed level of conversion weights for all population categories. This could occur normally because of response ratio and sampling biases. To illustrate the problem in the context of ABM, imagine that the male population of age 20 are weighted by 1:1 and the female population of age 20 are weighted by 1:2. Then, in the case of simulating a marriage event, each male will be married to 2 females. This result is not intended to be simulated nor reflect the actual weights in reality. Therefore, the weights should be adjusted to a constant ratio. In our case the original sample shows some fluctuation in the assigned weights. The average of all weights across districts and ages is 2.6% with a standard deviation of 4.7%. However, the median is 1.4% which is close to the predetermined 1.5% sampling weight. Therefore, the author believes that the sample should be adjusted to bring the weights to a minimal level of variation.

C. Data curation methods

The curation process for the households-based population is comprised of three key processes: temporal aggregation and missing data filters, boosting the sample representation level by demographic aggregation, and reweighting the sample. These challenges were addressed as follows:

C.1 Temporal aggregation and missing data filtration

For the temporal aspect, both time aggregation and data filtration were sorted at the same time. In order treat the data for these concerns, there are a few more matters to be considered:

1) The developed models in this research are based on five year time steps.
2) The households with no origin districts can be used as a training set to predict the allocation of newly migrated residents. This is based on the assumption that newly migrated residents do not answer the question of origin district in residential mobility. Therefore, these observations should not be fully excluded.
3) The complete exclusion of households who moved in year 2016 would mean that we lose a valuable 169 complete responses.
4) The old trends of residential mobility may not reflect the more recent trends. This is because new districts have been added to the city and old districts would lose their attractiveness over time.

Therefore, the suggested aggregation process includes the following:

1) The aggregation of data based on 5 years' time steps.
2) Exclude the observation before year 1991 or 2001
3) Exclude all incomplete surveys from year 2016.
4) Merge the 169 complete responses of 2016 to the group of data of time step 2010 to 2015. This is because of the marginal time difference between 2015 and 2016 and the small number of households in comparison to the total of the group.

See Table 31 for the temporal aggregation and filtration of ADA 2016 demographic sampling data

Table 31 the temporal aggregation and filtration of ADA 2016 demographic sampling data.

Year range description	Sample size	NA size		Complete responses
		Movement reason	Previous district	
1991 - 1995	**339**	**59**	**94**	**254**
1996 - 2000	**611**	**106**	**173**	**438**
2001 - 2005	**806**	**111**	**191**	**615**
2006 - 2010	**1517**	**142**	**317**	**1200**
2011 - 2016	**5043**	**2787**	**3050**	**1993**

Table 31 shows that the temporal aggregation and filtration process of the sample data reduces the sample size to 8316 participants with only 4500 complete responses. Moreover, the (2011 – 2016) time range included the largest number of samples which consists of 61% of all samples and 44% of complete responses.

C.2 Increasing the sample representation level using demographic aggregation

The classifications used for aggregating the data were as follows. For the age of the household's head, the subject specific classification of social stratification and mobility from the United National Provisional Guidelines on Standard International Age Classifications is used (UN, 1982). This classification grouped the age groups from 17 into 5 subcategories as follows: under 15, 15-24, 25,44, 45-64 and 65+. It is worth mentioning that the first category of under 15 was not included due to absence of household heads at these ages in the sample. Moreover, for the education level, the aggregation of its subcategories followed the 2011 UNESCO International Standard Classification of Education (ISCED) (UNESCO-ISCED, 2012). This classification aggregated the levels of education from 8 into 3 levels as low education, medium education and high education**.** Furthermore, for the aggregation of a household's size and the household head of work status, no international guideline for aggregation was found. It seems that the aggregation of these features is done on an ad-hoc basis. Thus, for the households' sizes, by and large the historical average of CFR was used as a threshold for aggregating the data. While any families with five children and above were considered in a single category, all the families of under five children were in separate categories according to their size as 0 child, 1 child, 2 children, 3 children and 4 children. The rationale behind this classification is that the age of the household head would increase with the number of children and therefore their mobility activity would decrease over time. Additionally, for the households' head work status. The number of subclassifications was reduced into two categories as employed and not employed for the sake of maximising aggregation. At this point, it is important to point that both household heads' marital status and the gender were not included. This is because females as household heads consist of a marginal number 4.4% of the total sample. Moreover, given the current large number of subclasses, marital status is temporarily ignored for the current analysis as it would increase the number of household's types. See Table 32 for a comparison of aggregated and disaggregated demographic subclassifications in addition to the source of reclassification guidance.

Table 32 a comparison of aggregated and disaggregated demographic subclassifications in addition to the source of reclassification guidance.

Feature	Categories reclassification		Aggregation guidance
	Old	**New**	
Age	**0-4**	Under 15	(UN 1982)
	5-9		
	10-14		
	15-19	15-24	
	20-24		
	25-29	25-44	
	30-34		
	35-39		
	40-44		
	45-49	45-64	
	50-54		
	55-59		
	60-64		
	65-69	65+	
	70-74		
	75-79		
	80+		
Education	**Literacy**	Low education	(UNESCO 2012)
	Reading only		
	Reading and write		
	Primary		
	Intermediate		
	Secondary	Medium education	
	University	High education	
	Postgraduate		
Households size	**0**	0	arbitrary
	1	1	
	2	2	
	3	3	
	4	4	
	5 - 20	5+	
Work status	**employed**	Employed	arbitrary
	Home wife	Not employed	
	retired		
	student		
	Unable to work		
	Unemployed not searching		
	Unemployed searching		

To examine the effect of aggregation shown in Table 32, we will compare the change in the values of the mean and standard deviation of four aspects: 1) the number of unique feeder zones to each zone; 2) the number of households moving from each zone to each zone; 3) the number of unique household types moving from each zone to each zone; and 4) the number of unique households moving to each zone. Moreover, the data will be compared for different scenarios as follows: 1) Data with complete records of mobility origin and destination; 2) data with no records on mobility origin; 3) the mobility zones at the district level; and 4) the mobility zones at the municipal level.

The justification for examining the exclusion of the residential mobility origins is due to the intention of modelling residential mobility of newly migrated households to Riyadh city which in this case has no original location. Also, it would give a better understanding of the effects of only considering the households characteristics to model their mobility choices.

Still, there are some limitations to be aware of in this analysis such as: 1) only the period between 2011 and 2016 will be tested. The change over the different 5-year time ranges from 1990 to 2015 will not be discussed. This is because the samples in these different time ranges are not from independent surveys and their sample size increases significantly over time as shown in Table 31. In the initial screening of the data, it was noticed that the variation of the aspects in question is positively correlated with the sample size. Thus, any observed changes over time will not be meaningful; 2) the comparison does not differentiate between the property ownership as rented or owned. All the moves are combined and treated similarly in terms of property ownership; and 3) Not all households' characteristics are included as mentioned earlier. Although all these excluded comparisons are possible to be carried out, the author believes that it is of less importance to the aim of this section given the current condition of the sample. See Table 33 for the aggregation effect on the representation level of the sample using demographically aggregated subclassification on two geographical levels.

Table 33 the aggregation effect on the representation level of the sample using demographically aggregated subclassification on two geographical levels.

Description	Value type	Districts (n=131)		Municipalities (n=17)	
		Disaggregated	Aggregated	Disaggregated	Aggregated
The number of all possible household types	count	19992	180	19992	180
The number of unique Household types in each sample	count	1200	114	1200	114
1- The number of unique feeding zones to each zone	Mean (% of total)	9.03 (6.89%)		10.12 (59.53%)	
	Std.	9.09		3.91	
2- The number of households moved from each zone to each zone	Mean (% of total)	17.08 (0.86%)		117.24 (5.88%)	
	Std.	26.6		132.14	
3- The number of unique household types moved from each zone to each zone	Mean (% of total)	1.90 (0.16%)	1.83 (1.61%)	10.49 (0.87%)	7.51 (6.59%)
	Std.	2.48	2.09	17.78	10.13
4- The number of unique households moved to each zone.	Mean (% of total)	21.35 (1.78%)	20.90 (18.33%)	160.7 (13.39%)	55 (48.25%)
	Std.	22.75	14.75	115.07	25.95

Table 33 shows that aggregating a household's characteristics significantly decreases the gap between the number of all possible household types and the actual number of unique household types available in the sample. While the total number of possible household types when using the four features of (age, education, size and work status) is 19,992 household types, the total available number of types in the sample data was 1200 types which covers only 6% of all possible types. After the aggregation process, the total number of possible household types decreased to 180 types. Moreover, the number of available household types in the sample became 114 types. Consequently, the data coverage of household types improved significantly from 6% to 63% with the demographic aggregation process. Moreover, when we compare the results of aggregating the households' characteristics and the geographic boundaries, we notice the following:

First, the change in the number of feeder zones changes only with the aggregation of geographical boundaries. As a result, the number of feeder districts (n=131) averaged 9 districts (SD= 9) which is about 7% of total districts in the sample. Alternatively, the number of feeder municipalities (n=17) averaged 10 municipalities (SD= 4) which is about 60% of total municipalities in the sample. Therefore, it can be argued that replacing or supporting the district's boundaries with the municipal boundaries would increase the geographic representation coverage of residential mobility trends about 10 fold from 7% to 60% in Riyadh city.

Second, the number of households moving from each zone to each zone is also changing with the aggregation of geographical boundaries only. Thus, at the districts level, the number of moving households (n=1993) averaged 17.8 (SD= 26.6) households. That is about 0.86% of the total sample size. However, at the municipal level the number of moving households (n=1993) averaged 117.24 (SD= 132) households which is about 5.88% of the total sample size. Although, it is expected to see an increase in the size of mobility with merging geographic boundaries, it is important to understand the empirical effect of such operations on the data.

Third, the number of unique household types moving from each zone to each zone is sensitive to the aggregation of households' characteristics and geographic boundaries. Therefore, when the household's characteristics are disaggregated. The number of unique households (n=1200) averaged at the district and municipal levels 1.9 (SD= 2.48) and 10.49 (SD = 17.78) households respectively. On the other hand, when the household's characteristics are aggregated, the number of unique households (n=114) averaged at the district and municipal level 1.8 (SD= 2.09) and 7.51 (SD = 10.13) households respectively, although the aggregation improved the sample representation of moving unique households by more than fourfold. At a maximum of 6.59%, all the averages are expressively low to represent the number of available unique household types when considering the origin zone of mobility, despite the substantial increase in the representation from 0.16 % to 6.59% with the aggregation process.

Fourth, the number of unique households moving to each zone is also sensitive to the aggregation of households' characteristics and geographic boundaries. Accordingly, when the household's characteristics are disaggregated, the number of unique households (n=1200) averaged at the district and municipal levels are 21.35 (SD= 22.75) and 160.7 (SD = 115.07) households respectively. However, when the household's characteristics are aggregated, the number of unique households (n= 114) averaged at the district and municipal level are 20.90 (SD= 14.75) and 55 (SD = 25.95) households respectively. It can be noticed that without considering the origin of residential mobility movement, the aggregation process becomes very influential for coverage level. The average sample representation of moved household types to each district improves significantly from 1.78% at the districts level with disaggregated household characteristics to 48.25% at the municipalities level with disaggregated household characteristics. This enhancement in the coverage level is believed to be important for the prediction of residential mobilities without their origin location.

Consequently, it is logical to think that the use of the original form of the data would not be acceptable due to the weak representation of the sample. However, aggregating the data attributes would increase the statistical representation of the data. Despite the fact that the aggregation process contradicts with the aim of this research as being district level oriented, and or for the sake of increasing the general accuracy of the developed model, it is believed that any prediction model of district level mobilities must include the municipal level as a first stage before advancing to the district level. Moreover, both origin and non-origin-based mobility prediction models should be tested to uncover the higher representation effects of prediction.

C.3 Reweighting the sample

The adjustment of the sample weights depends mainly on adding and removing respondents from the sample, This might occur while creating new observations that could be done by deriving techniques from the field of handling missing data such as Hot Deck Imputation, where missing values are filled from a pool of possible donors (Andridge and Little, 2010). Removing responses is also a common practice to deal with outliers or incomplete responses, although the use of such methods is normally done in a tight manner to avoid effects like creating biases the author tried to use these techniques in a systematic way to create an optimal sample out of the original sample. The steps to adjust the sample weights were as follows:

1) reducing the predetermined sample coverage level from 1.5% to 1%. In other words, this step should set the optimal sample size to 1%. Consequentially, removing sample respondents will be more likely. This approach is believed to be a less invasive than creating new respondents and will improve the model performance by reducing the number of active agents.
2) calculating the optimal sample size by converting the ADA population counts by district and age group into 1% population size that is rounded to the nearest 0.5.
3) comparing the actual sample size against the optimal sample size to determine the needed adjustment by removing or adding sampling participants.
4) calculating the possible change in the number of households for each district. This is done by dividing the total number of needed sampling adjustment in a district over a possible range of average household size. The possible average has upper and lower limits that are calculated from the original sample. This approach is used to create some sort of limits for changing the number of households as a controlling parameter to the adjustment process.

5) The process of adjusting the sample in each district is done by randomly adding or removing several households that are defined by the previous two steps, while the removal is limited by the household found in each district with no replacement of any existing households. The addition of households is open to the full list of households in the sample. Therefore, the current place of residence is changed to the new district. However, other aspects could be seriously affected by not being relevant to the observation such as the previous place of residence and the number of years lived in current property. This issue is one of the limitations that could be accepted on a random basis or excluded.

6) Step 5 is iterated over 200,000 times for each district and the best iteration is defined by the objective of minimising differences between the required change in sampling and the randomly selected household to be added or removed.

This calibration function was written in R and it can be described as a type of heuristic optimisation. The result of the adjustment process yielded a decrease in the average of all weights across districts and ages from 2.6% to 1.3%. Also, the standard deviation decreased from 4.7% to 1.4%. Moreover, the achieved median was 1.0% which matches the intended optimal ratio, and although, the average and the deviation rates are not perfect, they reflect an improvement in fluctuation reduction from the previous unadjusted sample. The method can be considered as promising in reweighting the sample but, it has shortcomings in how to define the number of households to be adjusted, the relativeness level of residential mobility features that are carried with each household, and the efficiency of optimisation process that must be further improved.

3.8.2 Built-Environment related data

The data for the built environment includes Riyadh city base maps, district attributes, existing schools, and building regulations. The ADA is the primary source for base map data, which is then merged with data from other sources, to produce a comprehensive dataset including all necessary components. The process of merging data is complicated due to the fact that distinct data sources have distinct naming conventions, geographical boundaries, and time reference. The following are the built-environment data in detail:

The building regulation data includes the standards for providing general education by stages and the Floor Area Ratio (FAR) set by the official planning body in Riyadh. The pre-designed school building prototypes, that belong to education data, are included here due to their contextual relevance to the provision standards of education.

Concerning provision standards, some upper and lower limitations are necessary for the calibration of the provision of schools, which represents the area designated for schools in land use planning. However, the primary challenge in determining these constraints is the diversity of school building and site components, such as the number of classrooms, internal and outdoor playfields, and parking spaces. Therefore, the provision guidelines from various sources are not directly comparable and determining the amount of space allotted to each student is contingent on the intended school's activities. Table 34 shows a summary of provision standards for general education by stage, gender and source.

Table 34 summary of provision standards for general education by stage, gender and source.

Source	Edu. Stage	Gender	Area per student (m^2)	
			Minimum	Maximum
(MOMRA, 2019)[1, 2]	Elementary	Both	4.40	
	Middle		4.70	
	Secondary		5.25	
	Multiple		4.80	
ADA 2016[3] (HCDA, 2016a)	Elementary	Male	1	
	Middle		0.65	
	Secondary		0.6	
	Elementary	Female	0.94	
	Middle		0.57	
	Secondary		0.5	
(Dudek, 2007)[3]	All	Both	1.8	3.5
(APS, 2021)[4,5]	Elementary		40	62
	Middle		54	67
	Secondary		76	87
(CDE, 2000)[3]	Elementary		59	86
	Middle		78	119
	Secondary		89	113
(DFE, 2014)	Elementary		2.9	4.5
	Middle		4.5	7.1
	Secondary		5	7.85
(GSE, 2020)[6]	All		15	22
(GOA, 2012)[4,5]	All		10	70

1: Minimum land area allocation per students for Riyadh city.

2: No mention of including parking area

3: The space allocated to the students is mentioned without specifics and in a broad manner.

4: A comprehensive guideline for the allocation of space that considers things like playgrounds, parking lots, and sports fields.

5: allocation of space based on: Ideal land parcel area/Ideal accommodated students.

6: allocation of space based on: constructed building footprint and outdoor area.

Table 34 demonstrates that the provision standards per students vary significantly across various sources. In addition, the criteria are usually subject to a minimum area allocation. Moreover, the ADA is the source of the lowest standards. Despite the observed diversity, it must be stated that there is some consensus among the various sources regarding some aspect such as the optimal school capacity for students by stage and the number of classes provided. The alternative approach to the area per student standard is the selection of pre-designed school types based on the availability of land. The MOE developed 22 building prototypes, each of which is either for a single educational stage or shared by several stages. Typically, either male or female students occupy each building. See Table 35 for the specification list of building prototypes developed MOE for education purposes.

Table 35 the specification list of building prototypes developed MOE for education purposes.

Edu. stage	Model number	Number of classes	Optimal capacity	Building footprint (m^2)	Area/student (m^2)
E	28.5	28	840	1003	1.2
E	21.5	21	630	1003	1.6
E	21.26	21	630	1797	2.9
E and M	12.23	12	360	554	1.5
E and M	16.24	16	480	740	1.5
E and M	3.3	30	900	1418	1.6
E and M	28.26	28	840	1715	2.0
E and M	28.2	28	840	1715	2.0
E and M	540.18	18	540	1418	2.6
E and M	9.24	9	270	740	2.7
E and M	33.11	33	990	5472	5.5
E, M and S	22.26	22	660	1702	2.6
M	26.5	26	780	1003	1.3
M	21.5	21	630	1003	1.6
M	15.5	15	450	1003	2.2
M	18.2	18	540	1418	2.6
S	24.5	24	720	1003	1.4
S	18.5	18	540	1003	1.9
S	26.2	26	780	1862	2.4
S	26.26	26	780	2072	2.7
S	20.26	20	600	1797	3.0
S	30.11	30	900	5472	6.1

E : elementary
M : Middle
S : secondary

Table 35 shows that the footprint area for the schools' prototypes ranges from 554 m^2 to 5472 m^2. Moreover, the area per student ratio ranges from 1.2 m^2 to 6.1 m^2 for the main buildings without its outdoor areas and other attached space. It is worth noting that the area per student adopted by the MOE are higher than those approved by ADA as indicated in Table 34. Moreover, the correlation between the prototypes footprints area and their capacity are not strictly linear.

In addition, as part of the building code, the current FAR is normally 1.2, with a maximum land coverage of 60 %. However, the latest proposed planning standards establish a maximum FAR of 3 for public services.

The data related to education, on the other hand, consists of the accessible white land parcels designated for education by the MOE. As of late 2018, 735 land parcels were indicated as being available, with available attributes including parcel ID, area, district name, and proposed usage by age and gender. The MOE also provided information about the current schools in the city of Riyadh. The data contains a list of 2486 schools and their characteristics. The attribute data includes latitude and longitude coordinates, the district name, the number of current students, the building capacity, the facility's suitability, the building's ownership, and the students' grade and gender.

In addition to the ADA base map data that covers numerous features at the parcel level, such as land use, area, and building status (i.e. built or not), other information was gathered, including properties prices, building permits and commercial activity permits.

The property price indicators were collected from the MOJ's open data website. The acquired data are derived from the Real estate indicator in neighbouring areas, which includes transaction data disaggregated by property type, value, area, and year. The obtained data cover the years 2010 to 2018 and were aggregated and averaged annually for residential lands in each district of Riyadh. Moreover, statistics on residential and commercial licences were collected from the Riyadh Municipality. The data primarily cover the years 2009 to 2018. The district-aggregated statistics included the number of issued residential licences, the total area of issued residential licences, the number of valid commercial licences, the area of commercial licences, and the number of distinct commercial activities.

3.8.2.1 Data quality evaluation

Due to the great degree of similarity in quality conditions amongst the various datasets, the quality of the data that is currently accessible will be examined in general, with a few highlights on the most important aspects.

Contextually, the data can be characterised as having been gathered for goals other than those of this research. For instance, school provision standards evolved much more from the perspective of functional school design than from the perspective of land use allocation in urban planning. There is a substantial quantity of information about land parcels, school characteristics, and property values. But there are significant challenges with the missing data problem. For instance, the location of privately owned schools is not included in the MOE data. Overgeneralization of indices as a result of data aggregation is another problem. For instance, in some areas, the price variance for the different local zones does not correspond to what the targeted population believes to be the average price.

In terms of timeliness of data, the collected data has the benefit of being based on records with time stamps, which solves the problem of data collection frequency to some extent. For the most part, the data accessible spans the years 2009 to 2018. Despite the fact that the historical depth of the data may not be sufficient to make projections with a high degree of certainty, it does give some insight into broad trends and is current in relation to the simulation base year of 2015.

The issue of inconsistency is one of the most significant concerns influencing the built environment data. This is due to the fact that the same feature may have multiple names or pronunciations across datasets, even ones that were provided by the same source. Another issue arises from the varied approaches taken to defining geographical boundaries. For instance, while the ADA counts some of the absorbed nearby provinces as being part of Riyadh city, other government agencies do not include them because they are outside of their administrative limits and do not consequently collect information for these regions.

In terms of intrinsic DQ, all data are provided from dedicated official bodies and are based on data records, which increases the reliability of the data. None of the data, however, includes notes about potential or known problems or a declared margin of error. Therefore, no data should be excluded from validity testing. The representational and accessibility of the data were limited by the coding of several crucial data, which rendered them illegible in their original form. The data was also delivered in digital formats and was simple to alter. However, in terms of accessibility, the data were made available upon request for the purpose of the study.

3.8.2.2 Data preparation, curation and future forecasting

This subsection provides a brief description of the problems that were discovered in the built-environment data as well as the solutions that were used to solve these problems. Following that will be a clarification of the overall method that was utilised to forecast the inputs that had a dynamic nature.

A. Data exploration and curation

A.1 Non-matching districts' names and boundaries among the different datasets

Due to the fact that the data at the district level were collected and subsequently combined into one set of data, there were issues with matching the naming and borders of the same features. In the meanwhile, in order to make the process of merging possible, the naming issue had to be manually unified across all of the separate datasets. It was necessary to aggregate the district boundaries according to the extent of disaggregation present in each of the data sets that were available for the same attribute. The end result produced 150 distinct districts, each of which had a unique level of data coverage due to the fact that the levels were determined by the final outcome. Figure 31 show a Riyadh city choropleth map for the level of data coverage for each district.

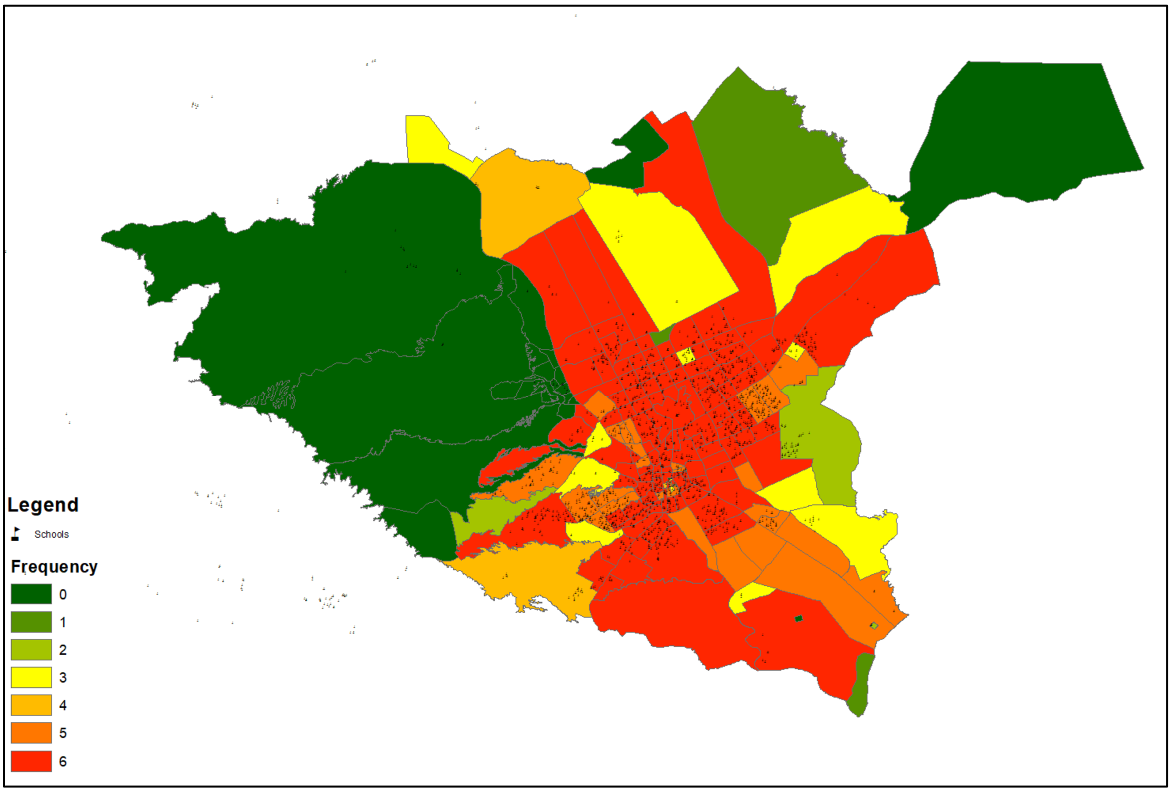


Figure 31 Riyadh city choropleth map for the level of data coverage for each district.

Figure 31 illustrates how the level of data coverage varies across districts, which can range anywhere from 0 to 6 distinct data sets. In addition, there are several smaller districts that are entirely contained within larger districts, and given that they have no data, it is possible that they may be dissolved. This example can be found in the outskirts, districts on the south side of the city. In addition, some of the locations of schools offered by the MOE are outside the boundaries established by the ADA for the city of Riyadh, which means they should be excluded. Additionally, on the northwestern side of the city, the enormous green periphery area that has no data is part of the governorate of Diriyah and not Riyadh city. As a result, there is no data accessible for the region, despite the fact that it is typically considered to be a part of the city of Riyadh.

A.2 Absence of private schools' locations

The missing locations for private schools were queried using the Google Maps API by school name, district names and education stage. Then, a sample of 100 schools were randomly chosen to manually check the result on Google Maps for location accuracy. The observed error level was less than 13%.

A.3 Average area of residential developments

Given that the area of available land for residential development is limited by the available amount of land and the propensity of annual development in each district, the number of issued residential building permissions was divided by the developed area to calculate the average area per building permission.

B. Future forecasting

The district-level data that could be utilised to simulate the residential mobility of individuals, such as the number of newly constructed residential units, the average residential land price, and crime rates, were extrapolated. During the development process, two major issues were identified that had to be addressed: the fluctuation in historical observations, which could lead to unintended sharp increasing or decreasing trends; and the avoidance of negative values for declining trends. While the former issue was resolved using a cumulative moving average for all previous data points, the latter issue was resolved using a power function that does not reach zero. Due to the lack of knowledge and reliance on the fitting goodness values obtained by the various extrapolation methods, these issues were initially overlooked. However, upon visualising the outcomes of the district-level model, it became evident that the generated patterns were not feasible for the near or distant future. Therefore, after investigating the cases of the strange outcomes and the identification of the early mentioned issues, it was concluded that the extrapolation of data must be thoroughly examined beyond the standard error indications and, eventually, at the individual level.

Chapter 4
Methodology

This chapter is divided into two main sections that address city-level and district-level population change respectively, along with their respective optimisation models. Before presenting these models in depth, the reasoning behind developing two separate models, an overview of the models with their inputs will be presented, and the shared methods for estimating population growth and the targeted school-aged population will then be explained. Afterwards, a description of the city level model and its optimisation problem will be discussed. At last, the district-level model will be described in accordance with the Overview, Design concepts, and Details (ODD) guidelines, along with its specific optimisation functions.

Two models were developed to support optimising the standards of providing education facilities using non-traditional methods. Although this research aims to optimise the land use provision standards at the district level, it acknowledges the difficulty of meeting this goal, which stems from the lack of data and planning support systems for this purpose. Therefore, a city level optimisation method for the standards was developed and applied to help in the case of lack of data particularly. The city level tool depends mainly on two key factors: a CCM population growth model and the amount of available land for education purposes, with the objective to balance the provision standards against the population growth, through the use of a controlling land consumption ratio throughout the time steps of the projection time horizon. However, when data is not an obstacle, the optimisation for the district level standards could be utilised. The district level model is similar conceptually to the city level in terms of accounting for population growth and education land availability. Nevertheless, the differences are significant in implementation as the district level model utilises an ABM instead of an accountant model and it thus has a household-based population structure instead of aggregated population counts. It simulates social change, predicts residential mobility, projects the district characteristics, and aims to balance the demand and supply by using pre-designed building units and possible specific land parcels instead of the area per person standard. It optimises for the whole projection time horizon at once instead of considering multiple time phases. Despite these differences both models can be conceptualised according to the following steps:

1) Population growth estimation
2) Locating the population at geographical boundaries (in the case of considering the district level)
3) Extracting the size of targeted population at predefined time horizons and,
4) Optimising the standards given targeted population, geographical boundaries, land availability and construction policies.

See Figure 32 for main modelling stages with its sub-models and inputs.

At this point, it is worth mentioning that the population growth estimation fundamentally follows CCM for both models that was described in Equation 8 for standard CCM formulation. However, in the following sub-sections the exact application for both modelling levels will be described in detail. Moreover, the targeted population in both models is the school-age population by education stage and gender. The different targeted populations are calculated using population synthesis of age-groups related to the students ages of each education stage. The synthesising method used was described in Equation 10 under the section of Data curation methods.

Figure 32 shows that fundamentally both main models consist of the same two main stages which can be labelled in a broad manner estimates of population change (at the targeted geographical level) and optimisation of standards. Therefore, the differences between the main models are related to the sub-models used and their methodologies. The natural increase in population is calculated by using the principles of CCM at both city and district level modelling. The change caused by residential mobility has three sub-models to choose from separately, which include Spatial interaction (SI), Neural Network (NN) and a Hybrid (Hy) approach that combines both methods. The number of inputs in the model of the district level categorically increased by over 50% when compared to the model of the city level which reflects the added complexity with disaggregated models. Moreover, these two main models are not linked programmatically despite the similarity of inputs, although the possibility of linking both models is standard and might lead to a more robust modelling practice. The comparable outcomes of these two models would be population estimates and the standards for providing facilities in the form of area per student. Obviously, this comparison would be subject to aggregation and averaging of outcomes. Each model will be explained separately in the following subsections.

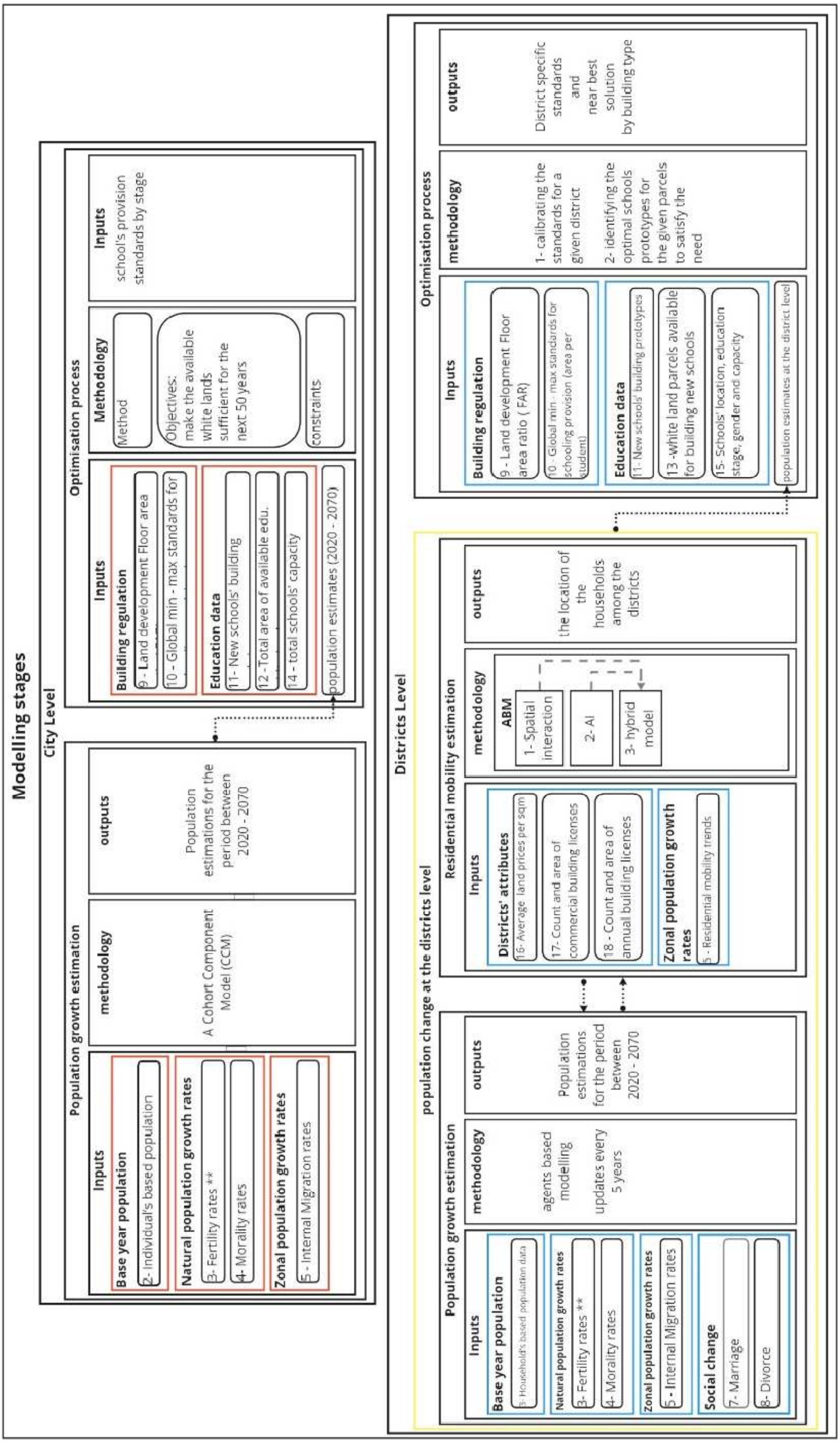


Figure 32 main modelling stages with its sub-models and inputs.

4.1 The city level model

The city level modelling can be viewed as the most aggregated attempt to solve the standards optimisation issue. The advantages of this approach lie in using commonly available aggregated data at the city level, the relative simplicity in used methods for population growth forecasting, and the expected increase in confidence levels and accuracy given the use of a fewer number of sub-models. Due to the aforementioned benefits and to evaluate the viability of the idea behind this research, this model was constructed as a baseline that holds the main components for standards optimisation to illustrate one of the proposed methods of facilities provision optimisation with minimal data required. The estimation method for population growth will be presented, and then the optimisation problem will be discussed.

4.1.1Population growth estimation

The CCM method for this model was coded in AnyLogic using the following sequence of events to reproduce the standard CCM method:

For the warmup phase, the model loads the base year population numbers along with the various rates of fertility, mortality, and migration. Then, the model iterate as follows:

1) The base year value is increased by 5 years.
2) Based on the new base year value and the selected growth scenario (i.e. high, mid and low), all demographic rates are updated. In this situation, fertility rates are updated depending on the year and the selected scenario, migration rates are updated based on the growth scenario alone, and mortality rates are held constant.
3) The base year population is adjusted based on migration.
4) For the female population we apply the following steps:
    a. The survived population numbers are calculated using the following equation:

$$P_{t+5}^{x,g} = P_t^{x-5,g} * Sr^{x-1,g} \qquad (16)$$

where:

$P_t^{x,g}$ Population P at age-group x and gender g in year t

$Sr^{x,g}$ The fixed survival rate Sr for age-group x and gender g

b. The new births numbers are calculated using the following equation:

$$P_{t+1}^{(0-4),g} = FP_t^{x_i} * Fr_{t+1}^{x_i,g} \tag{17}$$

which can be extended as:

$$P_{t+1}^{(0-4),g} = \{FP_t^{x_1,g} * Fr_{t+1}^{x_1,g} + FP_t^{x_{\dots},g} * Fr_{t+1}^{x_{\dots},g} + FP_t^{x_n,g} * Fr_{t+1}^{x_n,g}\} \tag{18}$$

Where:

$FP_t^{x_i,g}$ FP represent females potentially capable of childbearing in the list of age-groups x_i for a gender specific g new-born baby at year t

$P_{t+1}^{(0-4),g}$ The new-born population P annotated as age-group $(0-4)$ for gender g in year $t+1$

$Fr_{t+1}^{x_i,g}$ The specific Fr fertility ratio for age-group x_i to deliver baby with gender g in year $t+1$

5) For the male population, step 4.a (the calculation of survived population) is applied.
6) All population data are updated and stored.
7) The next iteration starts again from step 1.

After estimating the population change at the city level for the next 50 years in 5-year time steps, the expected number of enrolled students is approximated using the method described in Equation 10, which was used as part of the Data curation methods. At this stage the targeted population numbers should be ready to use as an input for the optimisation function which will be described below.

4.1.2 Optimisation at the city level

At the city level, the optimisation objective is defined to minimise the consumption of educational land serving the estimated number of students of coming generations by calibrating the area per students supply standards. Nonetheless, both the supply and demand sides of this optimisation issue contain various elements that must be explored and specified. This formalisation applies to the optimisation problem at both the city and district levels; however, in this subsection, the city level application is emphasized. The fundamental concept of optimising the provision of education standards can be expressed as follows:

$$needed\ education\ land_{t+n} < current\ available\ land \tag{19}$$

This assumes that current available lands are excessively offered. Therefore, in the opposite case, if the current amount of available land is below the minimum needed area while using the lower limit of global supply standards, the optimisation objective will be to minimise the amount of land acquisition necessary to satisfy the needed area. Therefore, if:

$$needed\ education\ land_{t+n}^{lower\ supply\ limit} > current\ available\ land \tag{20}$$

then, the optimisation objective becomes:

$$\begin{aligned} needed\ education\ land_{t+n}^{lower\ supply\ limit} \\ = current\ available\ land \\ + needed\ land\ area\ to\ be\ acquired \end{aligned} \tag{21}$$

Thus, Equation 19 and Equation 20 illustrate the two main cases of land availability for the optimisation process. To extend the main notion in Equation 19 we must consider the forming factors of both sides of the needed land area and the available land area. Firstly, at the side of needed land for education we consider student numbers, the provision standards (area per student) and the current capacity of students for the built schools, as shown in Equation 22 below:

$$\begin{aligned} needed\ education\ land_{t+n} \\ = (number\ of\ students_{t+n} - \text{current capacity}) \\ * provision\ standard \end{aligned} \tag{22}$$

In Equation 22 the term $(number\ of\ students_{t+n} - current\ capacity)$ can be shortened as:

$$surplus\ students_{t+n} = (number\ of\ students_{t+n} - current\ capacity) \tag{23}$$

Accordingly, Equation 22 can be also conceived as shown in Equation 24

$$\begin{aligned} needed\ education\ land_{t+n} \\ = surplus\ students_{t+n} * provision\ standard \end{aligned} \tag{24}$$

Equation 22 shows that the needed area is a product of subtracting the number of all students from the current capacity multiplied by the area factor per student (i.e. the provision standard).

Secondly, for the size of current land available, it is essential to consider the controlling factors (i.e. building regulations) imposed on the available land to calculate the net land area for development. These regulations include:

1. Floor Area Ratio (FAR), and
2. Maximum Building Footprint Area Ratio (MBFAR).

These factors can be formulated as follows in Equation 25:

$$net\ available\ land\ area\ for\ development = current\ available\ land * (1 - MBFAR + FAR) \tag{25}$$

Where:

FAR	A fixed ratio of 120%
$MBFAR$	A fixed ratio of 60%. This policy does not normally include the playground area for sports. Therefore, the remaining area of 40% is included to account for this area if needed.

Equation 25 shows that actual land area can be different from the net development area. Moreover, the actual land area and its factors can be conceived as an approximation of net development area. Therefore, the usable area of any land is estimated to be at 160%. Another crucial element that would have to be included in the formula to account for any problems arising from underestimating the required area is the proportion of land that must be left undeveloped. Unfortunately, if the amount of available land is typically little, this aspect is expected to drive the optimisation process towards a lower bound of standards. Equation 26 shows the addition of the land preservation ratio in determining the net available land area for school development.

$$net\ available\ land\ area\ for\ development = [current\ available\ land * (1 - MBFAR + FAR)] * land\ presertaion\ ratio \tag{26}$$

Therefore, joining both the needed area and the available area sides of the optimisation problem can be written as:

$$(number\ of\ students_{t+n} - \text{current capacity}) * provision\ standard < [current\ available\ land * (1 - MBFAR + FAR)] * land\ preservation\ ratio \tag{27}$$

To simplify Equation 27, and reformulate the optimisation concept in Equation 19, it can be shortened and expressed as:

$$surplus\ students_{t+n} * provision\ standard < net\ available\ land\ area\ for\ development * land\ preservation\ ratio \tag{28}$$

Given the annotation in Equation 28 it can be said that the optimisation of education standards is concerned with allocating the surplus of students in a futures perspective within a limited amount of land and known practical limits of required area per student. However, Equation 28 considers the whole projection period and it can be disaggregated by, for example, 5-year time steps. The disaggregation of the optimisation process could then be applied to the net available area and the land preservation ratio, which will allow us to time the response of schools provision accordingly – see Equation 29.

$$\begin{aligned} &surplus\ students_{t+n} * provision\ standard \\ &\qquad < net\ available\ land\ area\ for\ development_{t+n} \\ &\qquad * land\ preservation\ ratio_{t+n} \end{aligned} \tag{29}$$

By using this strategy, only one provision standard per educational level will be optimised over the entire time horizon, but each iteration will also account for the number of surplus students and the ratio of land preservation. In the long run, this strategy will guarantee that the available land will not be completely consumed. Additionally, each stage should maintain a single standard for an extended length of time. Therefore, the city level optimisation objective can be written as:

$$\min_{\substack{s.t.\\ p\in[0.001,0.7],\ \ 0<p<0.7\\ x\in[ele,mid,sec]\\ d^{ele}\in[1,3.9],\ \ 1<d^{ele}<3.9\\ d^{mid}\in[0.8,2.7],\ \ 0.57<d^{mid}<2.7\\ d^{sec}\in[0.6,1.7],\ \ 0.5<d^{sec}<1.7}} \sum_{t=1}^{10} abs[(l_t\, p_t) - s_t^x * d^x] \tag{30}$$

Where:

l_t	Total available land at time t.
p_t	land preservation ratio at time t, subject to constraints between 0 and 0.7.
s_t^x	The students surplus at time t and education stage x for both genders.
d^x	The calibrated provision standard for education x.
ele	Elementary stage.
mid	Middle stage.
sec	Secondary stage.

Equation 30 shows that for each time iteration, the land preservation ratio ranges from zero to 70%. Additionally, distinct maximum and lower constraints for the various educational levels were introduced, and the lower bound limit was set to match the ADA's current guidelines.

The proposed optimisation function will iteratively ensure that the supply side represented by the inverse of the land preservation ratio is equal to the demand side generated by the number of surplus students for every time step through the targeted time horizon. However, it may be claimed that one of the shortcomings of this optimisation strategy is that it does not account for the lowering of currently offered school spaces to match instances of declining demand.

4.2 The districts level model

First, the ABM will be described in accordance with the ODD protocol for the district level modelling. Then, a description of the optimisation function will follow.

4.2.1 Purpose

This model illustrates the possibility of developing urban planning regulations using advanced approaches instead of the common practice of arbitrary global averaging of standards in urban planning policies. This model is developed for the purpose of optimising the supply standards for schools according to estimated school age population at the district level for a targeted year, within defined limitations such as land availability and architectural building standards. To achieve the model's purpose, the model adapts known and novel mechanisms for simulating population growth using CCM, applying social changes based on logically occurring probabilities, and predicting household residential mobility using traditional SI modelling or in the less common practice using NN models to estimate the school-age population within each district. Finally, a district specific optimisation process is established to optimise the standards of schools' provision.

4.2.2 Entities, state variables and scale "Ontology"

4.2.2.1 Entities

A. Environment and Spatial units

Time is represented in a discrete time steps, each of which accounts for 5 years. The main physical environment is a continuous space that is drawn based on the geographical boundaries of the districts of Riyadh. Agents utilise a metric unit for distance with a granularity of 1 meter. The districts' boundaries were imported from ADA GIS base maps for Riyadh city, producing 150 districts; there exist also 2 additional areas for holding agents temporarily during allocation events and for storing agents after their life cycle. The administrative boundaries of the city's districts are the main space for the agents to move within. The location of agents is randomly assigned within each district. Therefore, the household containers and agents are scattered within the administrative boundaries of their current district. See Figure 33 which is the visual representation of administrative boundaries of Riyadh districts in the model interface.

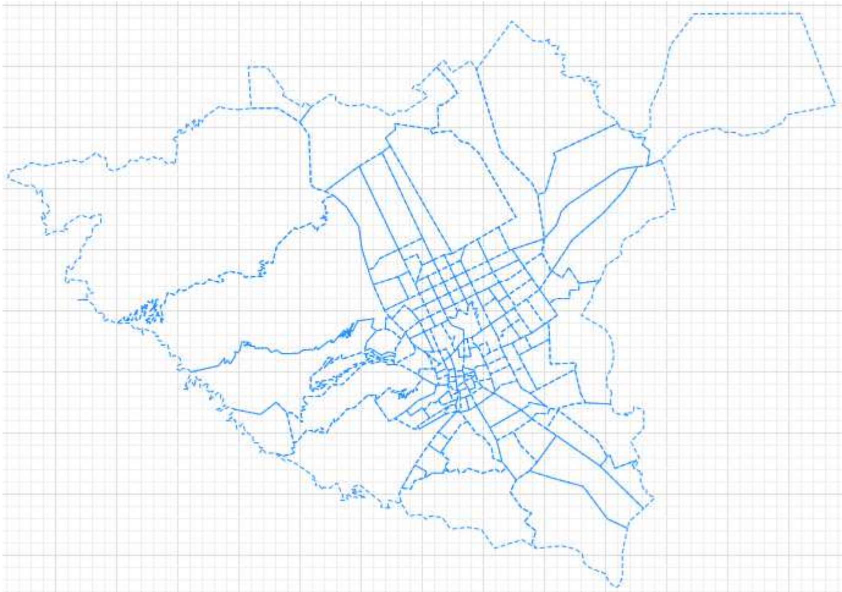


Figure 33 the visual representation of administrative boundaries of Riyadh districts in the model interface.

Figure 33 shows the representation of Riyadh city's 150 administrative boundaries (i.e. districts). These boundaries have no overlapping areas, and any continuous lines are produced by overlapping dashed edges.

B. Individual agents

The model has two types of individual agents: individual people and city districts as agencies. Individual people are tied to a collective agent through their attributes to form a household unit.

First, people ("population") agents have different functions and connections based on their gender and the type of relationship they have to their household container (i.e. their collective household agent). Therefore, the individual population agents were distinguished by creating four subtypes: Father, Mother, Male child and Female child. Second, each District agent is characterised by 14 attributes, of which 8 relate to education status. Current schooling capacity is disaggregated by gender and stage. See Table 36 for all types of individual agents in the city level model with its attributes and functions.

Table 36 shows that population agents are characterised by 8 attributes, 5 of which are demographic attributes, and 3 of which relate to residential mobility. The Death function is the only shared function among the subtypes of population agents. Both the Birth and

Propose marriage functions are subtype-specific functions. This is because agents of type mother are the main element of population reproduction. Likewise, Male children will start proposing to possible spouses at a certain time during their life cycle. The types of connections a population agent can have to their associated household agent are defined in the collective agent's section below.

Table 36 all types of individual agents in the city level model with its attributes and functions.

<table>
<tr><th>Distinctive agent type</th><th>Shared attributes</th><th>Shared functions</th><th>Distinctive functions</th></tr>
<tr><td>Father</td><td rowspan="4">Status attributes
a. Age (int.)
b. Income (int.)
c. Employment (text)
d. Education (text)
e. Removed agent. (true/false)
Location attributes
f. Willing to move from current home (true/false)
g. New home location (0-150)
h. Release from temporary location (true/false)
Social attributes
i. links to other agents</td><td rowspan="4">death</td><td></td></tr>
<tr><td>Mother</td><td>Birth</td></tr>
<tr><td>Male child</td><td>Propose marriage</td></tr>
<tr><td>Female child</td><td></td></tr>
<tr><td>Districts</td><td>Status attributes
a. Land Price (int.)
b. Number of unique commercial licences (int.)
c. Count of current residential building licences (int.)
d. Distance to all other districts (dubl.)
e. Capacity limit of residential licenses (true/false)
f. Number of issued residential licenses (forecasted int.)
Education status attributes
g. List of all white land parcels. (int.)
h. Total area of available white land parcels. (int.)
i. Current capacity of schools by stage and gender (int.)</td><td></td><td></td></tr>
</table>

C. Collective agents

The developed model uses a collective agent that is called a Household to tie all individual agents of the same family in order to form the main modelling household unit. This household unit can also be called the household container. The use of this container agent is not only to link the individuals of a family together but to represent the decision-making unit with its characteristics. This decision-making unit is an idealised image of the family that is limited to cultural norms in the Saudi context. Therefore, aspects such as the independence of adult children without marriage and family participation in decision-making are other aspects to be considered in future developments. Also, the household agent handles the possible social changes among the family members resulting from events such as one or both parents' death, divorce, remarriage and changes in children's custody. The household agent links the 4 types of individual population agents (Father, Mother, Male and Female child) as family members, and one additional link to the current district. Figure 34 for the main structure of the collective agent (household) with its links to individual agents.

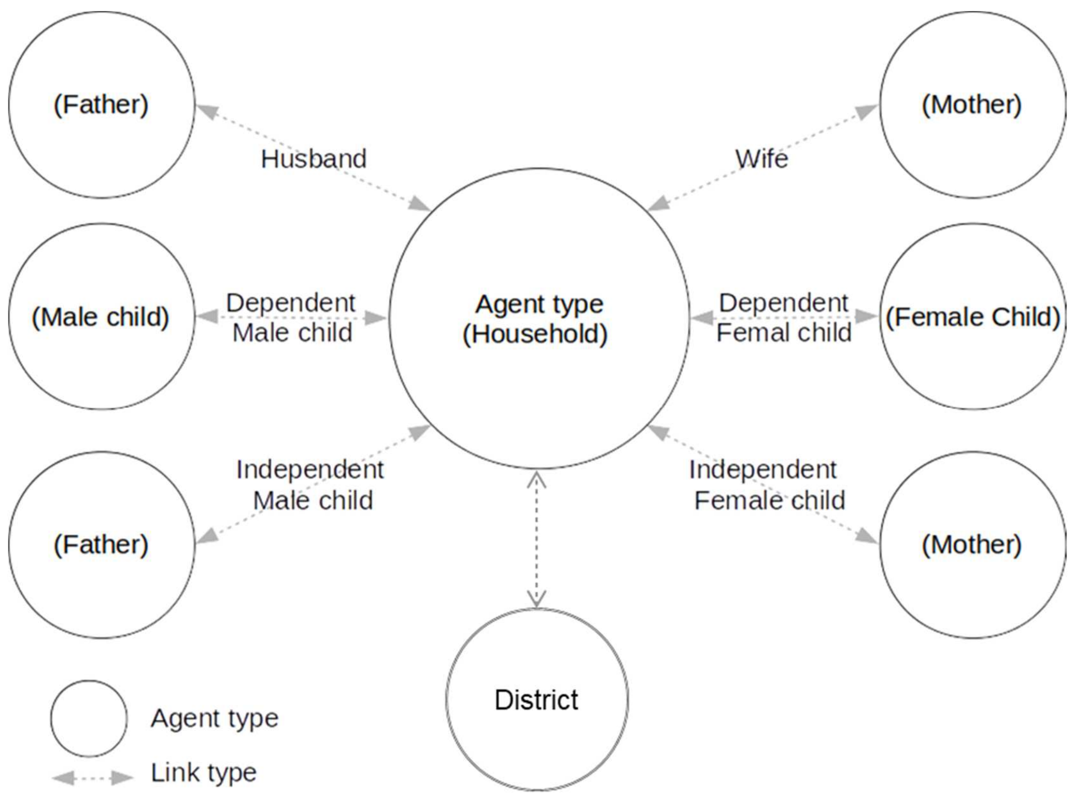


Figure 34 the main structure of the collective agent (household) with its links to individual agents.

Figure 34 shows that the household agent is the core that links all individual agents together. The main difference between dependent and independent children is their marital status. This is because agents who were formerly children but now of type Father or Mother will maintain a tie as an independent child to the household they grow up in, even after establishing a new household, for possible child custody assignment reasons and record tracking.

The household agent contains its own attributes and functions. First, the attributes which resemble the characteristics of the household head includes all the shared attributes of individual population agents mentioned in Table 36 (i.e. age, income, current district, etc.), in addition to several unique attributes. For a detailed overview of the household agent attributes see Table 37 for the shared and unique attributes of the household agent.

Table 37 the shared and unique attributes of the household agent.

Aspect	Shared attributes with individual agents	Unique attributes to the household agent
Status attributes	**Drawn from head of HH:** a. Age *(int.)* b. Income *(int.)* c. Employment *(text)* *d.* Education *(text)* e. Removed agent. *(true/false)*	a. newly created HH *(true/false)* b. HH head exists *(true/false)* c. total number of children *(int.)* d. HH head willingness to remarry *(true/false)*
Location attributes	a. Willing to move from current home *(true/false)* b. New home location *(0-150)* **c.** Release from temporary location *(true/false)*	a. Initial home location *(0-150)* b. The spatial allocation cost of current location *(int.)* c. Estimated duration of living in current district *(int.)* d. *Distance to all other districts* e. Home ownership *(true/false)*

Table 37 shows the attributes for the household agent, including variables and parameters that are mostly for checking the agent condition in order to trigger related functions such as who is the household head. Home ownership could be either rented or owned categories, while the values of shared status attributes in the household agent are updated using the household head characteristics, and the location attributes of the individual agents are updated using household agent values.

The household structure is one of the most distinctive characteristics of this model, as it reflects the actual relationships between family members and their place of residence, establishes the decision-making unit, and enables the traceability of family members as they become independent. Although the traceability of family members is not used in the current context, it could be used in the future when considering complex decisions including residential relocation near family members. We now present the modelling overview.

4.2.2.2 Process overview and scheduling

As the model steps forward in time, the following processes take place in the given order:

1. Update the environment.
    a. Increase the year date by increasing 5 years.
    b. Decrease the amount of white land for residential purposes in every district using the forecasted number of issued residential building licenses if possible.
2. Update demographic and social status of individual agents.
    a. Increase the age variable of all individual agents by 5 years.
    b. Check the death probability of all agents: disconnect those who died from their household containers and move dead agents to the cemetery/storage location.
    c. Check the marriage probability for all adult male children.
        i. If any candidates for marriage, search for a female adult willing to marry.
        ii. Establish a new household agent container and link marrying agents to the newly established household agent.
        iii. Assign the household head attributes.
        iv. Assign the household's district location, drawing from either:
            1. A uniform distribution probability for the 150 districts.
            2. The NN probabilities generated using households with no previous location.
        v. Move the household and marrying agents to their new location and estimate the duration the newly created household will remain in its initial location based on the selected residential mobility method.
        vi. Check the birth probability of all mother agents. Birth probability is conditioned by mother's age, birth's gender and availability of connected husband to the household.
        vii. Create new male and female children based on their probabilities.
        viii. Link the new agents to the household agent container and move the agents to the household district location.
3. Update the demographic and social status of household agents with X probability.
    a. Divorce (if both parents are alive)
        i. Create a new household container for the separating agent.
        ii. Keep the children with their mother or move them into their father household agent based on custody probability.
        iii. Assign the separating household's district location, drawing from either:

1. A uniform distribution probability for the 150 districts.
2. The NN probabilities generated using households with no previous location.

iv. Move the newly created household and its individual agents to the new household's district location, then estimate the duration the newly created household will remain in its initial location based on the selected residential mobility method.

b. Children custody (if both parents dead)
   i. Check for availability of adult dependent sibling to be the household head. If not possible then:
   ii. Check for availability of independent sibling to move all siblings into their household. If not possible then:
   iii. Pick a random household to take all children under their custody.
c. Update the household head based on randomly weighted probability
   i. With some probability, all father characteristics will be assigned to the household head.
   ii. If mother probability is true then, all mother characteristics will be assigned to the household head.
   iii. Otherwise, all father characteristics will be used for the household head except for the income parameter which will be summed for both parents.
d. Marrying in case of a single parent: if single parent is willing to marry, then:
   i. Look for a random widow/widower willing to marry (conditioned by suitable age difference). If found then,
      1. Disconnect the found agent (fiancée or fiancé) from its current household container and connect it to the proposing agent household (i.e. suitor household or female suitor household)
      2. Update the fiancée or fiancé agent location attributes and move them to their new household's district location.
      3. Disconnect the fiancée or fiancé children from their current household and connect them to the proposing agent household.
      4. Update all children associated with the fiancé/e location attributes.
      5. Set the decision maker for the household head to be the father agent.
      6. Remove the fiancée or fiancé household container to the storage location.
   ii. If no widow or widower are available then, look for a non-married adult agent from the children's classes. If the suitable agent is found then cast the child agent into a father or mother by:
      1. Add a new father or mother agent.
      2. Update the new agent status and location attributes.
      3. Connect the new agent to the proposing agent household.
      4. Remove the child agent to the storage location.
      5. Set the father agent to be the household head.

4. Remove all inactive household agents to the storage location.
5. Update all household head ages.
6. Add migrated population.
    a. Define the number of population individuals to be added by the current iteration.
    b. Each migrant will be added with its household container. Therefore, the process is as follows:
        i. Create a new household container.
        ii. Randomly assign the household head age following a weighted probability.
        iii. Randomly assign the household head gender using a weighted probability.
        iv. Assign the household's district location, drawing from either:
            1. A uniform distribution probability for the 150 districts.
            2. The NN probabilities generated using households with no previous location.
        v. Based on the household's head gender, age and location, create the migrated individual agent as father or mother agent.
        vi. Connect the individual agent to its household container and move the individual to the household's district location.
        vii. Estimate the duration each newly created household will remain in this location based on the selected residential mobility method.
7. Apply residential mobility modelling on households.
    a. Log each household current location and year.
    b. For every household with estimated house living duration greater than zero years, decrease the living duration by 5 years.
    c. For every household whose home living duration reached zero, run the residential mobility function which has three alternatives to choose from as:
        i. The SI model
            1. Choose if the ratio of moving households is manually capped or a result of the mobility threshold parameter in spatial cost comparisons.
            2. Calculate the spatial cost of the current location based on the household attributes (age and number of kids) if identified within the calibrated types or based on the district attributes only otherwise. This step includes check the amount of white land available for issuing residential building licenses.
            3. Compare the current district cost with all other districts.
            4. Shortlist the possible districts to move to using a defined threshold value.
            5. If the list of possible districts has candidate districts, then each candidate will have a weighted probability based on its cost.
            6. If the list of possible districts is empty then, the household will remain in its location.

    ii. The NN model
        1. Find the precalculated probability of residential mobility by household type (age, kids, and location).
        2. If the location is not calibrated, find the precalculated probability of residential mobility by household type (age and kids).
        3. All districts are considered for residential mobility based on weighted probability.
        4. Find the precalculated probability of home living duration by household type (age, kids, and location or age and location).
        5. All durations are considered for home living duration based on weighted probability.
    iii. A hybrid model that combines SI and NN
        1. From the (SI) model apply steps 2, 3, 4, 5 and 6 to estimate the weighted probability of districts for the residential mobility move.
        2. From the (NN) model apply steps 4 and 5 to find and assign the home living duration.

8. Finalise the iteration by required collecting data.
    a. Estimate the population numbers by:
        i. Convert the population sample into population counts.
    b. Estimate the schooling age population.

The processes scheduling can be seen to reflect a strong hierarchy of logical dependency among the model functions. For example, the birth function comes after aging, death and marriage functions. However, the effect of social change and migrated population are added afterwards but could be rearranged to be before the birth function. At this stage the model imitates the Cohort Component Modelling in estimating the population growth which normally adds the migrated population at the end of an iteration. Also, our model does not impose the effect of social change (at least at the first iteration) in birth estimations. However, the order of events can be altered to include these effects of social change and migration.

4.2.2.3 Modelling concepts

A. Basic principles

The modelling approach for the districts level is fundamentally based on individual household decisions. Therefore, the common practice methods for estimating population growth and residential mobility were adapted to work within the ABM environment.

B. Emergence

The total number of pupils in each district emerges from the population reproduction process and the residential mobility trends. The reproduction process itself is influenced by fertility rates, survival rates, migration and social changes, residential mobility can be imposed with historical trends captured using NN techniques or emergent from changes in the districts' attributes using SI modelling.

C. Adaptation

Relocation is an adaptive behaviour employed by household agents. Households decide where to move to and how long to stay in a location. The model user can choose between two main decision functions for the relocation behaviour which can be classified as direct objective seeking for SI and indirect objective seeking for NN.

D. Objectives

Household agents using SI have a direct objective seeking behaviour of relocation, each agent tries to maximize its utility function, comparing what the districts can offer in terms of residential units' prices, availability of new dwellings, variety of shops and distance to its current location/community. All these factors are assessed differently by the different household types (i.e. households with different attributes) with the model calibrated for its current location, household head age and family size. It must be noted that in the case of a household type not being available among the calibration sample, there are general calibration weights based on the whole sample to be used alternatively. This objective is measured using Equation 31.

$$P_{ij}^{t} = \frac{e^{(l_j.x_l^t)+(s_j.x_s^t)-(p_j.x_p^t)-\left(d_{ij}^{\beta}.x_d^t\right)\, a^t}}{\sum_{ij} e^{(l_j.x_l^t)+(s_j.x_s^t)-(p_j.x_p^t)-\left(d_{ij}^{\beta}.x_d^t\right)\, a^t}} \tag{31}$$

Where:

P_{ij} is the probability to move from district i to j for households of type t
e is Euler's Number (2.71828), associated with the natural exponential function
l_j is the number of issued building licenses in district j
x_l is the weight of issued building licenses for households of type t
s_j is the count of unique commercial licenses in district j
x_s is the weight of unique commercial licenses for households of type t
p_j is the average residential white land prices in district j
x_p is the weight of average land prices for households of type t
d_{ij}^{β} is the distance between district i to j with an exponential friction factoring β
x_d is the weight distance for households of type t
a^t is a scaling constant for households of type t

Equation 31 hypothesises that higher prices and distances would have a negative effect on the expected utility. On the contrary the number of new building licenses and diversity of shops are anticipated to have a positive effect on the utility cost. Moreover, distances between districts are modelled with an exponential effect.

E. Sensing

In residential mobility all households are assumed to know the changes that occur in the districts' attributes and calculate its utility cost; accordingly; they have perfect information. This assumption is obviously not entirely true in reality, and more nuanced sensing could be introduced in the future development of this model. This behaviour is embedded within the residential mobility choice, in which the agents consider district prices, commercial activity, distance from current residence, and the number of new residential units.

F. Interaction

A number of direct interactions exist among agents. Birth events within agents of type Mother creates a new baby agent that is directly connected to its mother and household. However, agents of type Mother cannot reproduce without the mediation of a household agent that links it indirectly to a father agent. Household formation is another case of direct interaction: if both agents were previously unmarried, a new collective agent of type Household will be created, and each agent will be cast depending on its gender as either father or mother before being connected directly to this new mediating Household agent. Also, these married agents will disconnect from their original household as dependent children to be reconnected as independent children. In cases of remarriage, the candidate partner will move to the suitor's household agent alone if they were not married before, or with its children if they exist. Third, in the case of divorce, based on some probability one of the partners will establish a new mediation agent of type household instead of returning to its original household. This is because this event will affect the children which might move with the separating agent. Fourth, in cases where both parents are dead and dependent children have become independent, the household agent will be removed out of the active agents list. See Figure 35 for the life cycle of households within the district level modelling for an illustration of agents and connections.

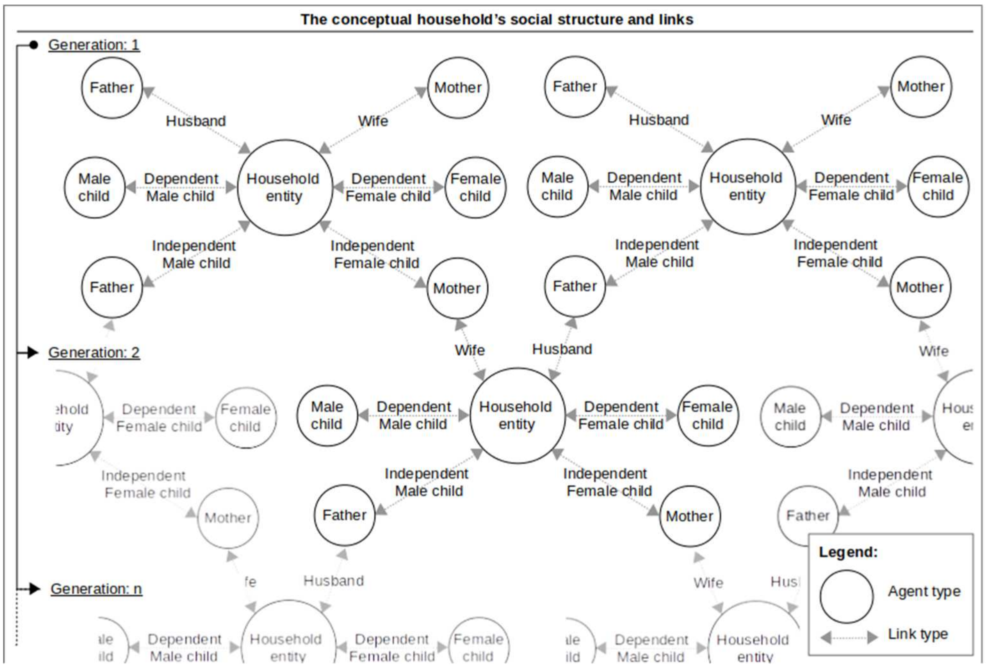


Figure 35 the life cycle of households within the district level modelling.

Figure 35 shows the evolution of population generations through Households and the types of connection each agent has.

In models of residential mobilities driven by district characteristics, indirect model interactions can be detected. For instance, the number of new building permits will continue to consume the whiteland area until there is no more space for additional development. Therefore, the attractiveness of the new buildings in some districts will impact the types of newcomers. This type of indirect effect is applicable to all district characteristics considered by mobility models, such as land prices, etc.

G. Stochasticity

Stochasticity plays an important role in the model for most aspects of simulation modelling. In population change, all births, deaths, marriages, divorces, child custody and defining the household head rates are stochastic within specified frequencies. Moreover, the number of migrated individuals has been randomised by a range of (+-/10%) to remain within a known range but, without having to model the variation cause. Furthermore, because some rounding of population numbers was made during the adjustment of main population sample, the counts rounding could be inverted by adding a random fraction that (+-/10%) of individuals for each district. In addition, unlike residential mobility that takes a deterministic approach in selecting the most likely alternative among the districts, this model adds a stochastic element by considering a weighted probability for all the selected alternatives. While districts attributes are not dynamically created, it is possible to include such a feature in future developments.

H. Observations

The user can see an interactive list of occurring main events with text that describe the year, the main modelling stage and the sub modelling stages. Geographically, the user can monitor the movement of agents among the districts of Riyadh city, which includes the movement between temporary and actual locations. In addition, the connections linking the household agents with their individuals can be enabled by drawing lines connecting the agents together. These drawn lines are updated simultaneously when events change. Also, the storage locations for agents with their finished life cycle are shown. These representations enable the user to have a better idea of changes occurring in terms of population demographics and mobility trends. The statistical data is collected at the end of each iteration, and it includes general information about the population that covers the number of agents by entity type. Also, population pyramid charts are provided. The targeted data of school age population is also collected and stored in separate lists to be accessed later by the optimisation function.

4.2.2.4 Initialisation

For the population initialisation, the population of collective household agents consists of 7026 agents. These agents are created based on the adjusted 2016 ADA demographic survey data, which also contains the numbers of individual agents to be created and linked to their collective household agents. Moreover, the data includes all default values for the variables of age, education, home ownership, employment, income, household head and location. These imposed values will ensure that all agents are placed in their initial locations and wait for update. It is important to know that the value of home living duration is set to be (0) to ensure all agents will call the SI model directly when choosing the SI model for residential mobility or being promptly estimated with the choice of NN model. This aspect leads us to the different scenarios the user is expected to choose from at the initialisation stage.

For the possible scenarios, the model user can choose from three methods for residential mobility which are SI modelling, NN modelling and Hybrid modelling that combines the SI and NN approaches. These options do not affect the initialization setup of the model, they only change the residential mobility function when called.

For the environment initialisation, the model is fed with a GIS base map for Riyadh city that was originally sourced from ADA. The base map consists of 150 districts presented as polygons. Moreover, four additional spaces were created to handle temporary holding of agents while in relocation and in the permanent holdings of unactive agents. Each of these polygons then, was linked to an individual agent as a type of district, the values of the districts' variables being fed using separate databases.

4.2.2.5 Input data

See Chapter 3 for a detailed explanation of all data sources, data classification, data timeliness, data quality evaluation, concerning issues, data curation and future estimations to drive the different simulation processes. Also, see Figure 32 for the required input data by modelling stage.

4.2.2.6 Sub models

Several sub models were created to for the district level ABM which consist of the following:

A. The sub models of residential mobility

Three main models are used which include SI, NN and hybrid.

A.1 The sub-model of SI modelling

The SI model is one of the methods adopted by this model to simulate and estimate residential mobility. SI provides the means to estimate the probability of a household to move to all available districts based on their spatial cost. Actual moves can then be triggered accordingly if the spatial cost exceeds a certain threshold of benefit gain. This method was calibrated based on disaggregated household's attributes besides the district's attributes. It is worth mentioning that the household's attributes of age and the number of children were uniquely coded by concatenating their values as a string. This coding is used to look and find the associated values of each household type. However, due to limited coverage of household types in the sampling data, a generic household type is calibrated to substitute the absence of household type specific calibrations. The SI model executes as follows:

First, for any household agent entering the SI model, check if it has calibrated weights for the household attributes of age and count of children. If the household type is calibrated then, get its weights values or get the generic household type values otherwise. These weights include the threshold value which control its move. Then, we get the attributes of price, distance, commercial and residential licenses for all districts and normalize them using their max values individually. The reason for normalizing the districts attributes inside this step is to simplify the steps required in changing the input data. Afterwards, apply Equation 31 the SI equation for calculating the household utility function for all districts. After that, filter the districts based on comparing their utility cost with the current district cost plus some benefit margin. See Equation 31 the SI threshold to filter the candidate district moving to and to trigger the move.

$$Cd = c_j \geq c_i * threshold \tag{32}$$

Where:

Cd	the list of candidate districts
c_j	the estimated cost for the destination district j
c_i	the estimated cost for the origin district i
$threshold$	the threshold value to consider the candidate district j

After filtering candidate districts, if the number of candidates is one or higher, the movement trigger parameter will be set as true. Also, the district's probability list will be updated with the probability of districts that exceeded the benefit threshold. Otherwise, the movement trigger will remain false and the district probability list will not be updated.

In calibrating the SI model, the population sampling data was split as training data covering the period between (2006 – 2010) and prediction data covering the years (2011 – 2015). Moreover, all population growth and social change events were disabled to stop the population demography from changing. Also, the modelling time was limited to one iteration for both training and prediction runs. Due to the sampling nature, the types of unique households differ among training and prediction data in terms of attributes and counts. The number of unique households in training and prediction data were 19 and 20 types consecutively. Therefore, the disaggregated calibration of households does not cover a large amount of household types and require a generic household type to cover the lack of household types.

Before showing the calibration results, it is important to understand the expected effects of the district's attributes that are part of the SI mechanism. The correlation between the residential mobility inflows was tested in both training and prediction data against the count of issued residential building licenses, the average square meter price for residential land, and the count of unique commercial licenses (i.e. activities) at the district level. See

Table 38 for the Pearson correlation coefficient test for residential mobility flows against the district's attributes.

Table 38 Pearson correlation coefficient test for residential mobility flows against the district's attributes.

Variable	Year	Inflow 2010	Inflow 2015
Count of issued residential building licenses	2009	0.90	0.77
	2010	0.72	0.92
	2014	0.59	0.80
	2015	0.53	0.73
Average square meter price for residential land	2010	0.05	-0.03
	2015	0.09	0.05
Count of unique commercial activities	2010	0.15	0.13
	2015	0.34	0.30

Table 38 shows that only the number of issued residential building licenses are significantly correlated with the inflow of residential mobilities largely averaging (r =0.75). Also, we notice that there might be a delay effect between the time of issued licenses and residential mobility. This is because residential licences of previous years are more correlated to inflows of residential mobilities than licenses issued in the same years – after all, residents cannot move in until after the new buildings have been built. The correlation between residential licences in years 2009 and 2010 to residential mobilities in 2010 were (r= 0.9) and (r= 0.72) consecutively. Similarly, the correlation between residential licenses in years 2010, 2014 and 2015 to residential mobilities in 2015 were (r= 0.92), (r= 0.8) and (r= 0.73) successively. This delay effect is expected as the duration between getting the building permit and the actual move of households would normally take more than a year. On the other hand, both average land prices and the number of commercial activities has no significant correlation to the inflows of residential mobilities. It can be said that residential building licences seems to be a principal component in predicting mobilities for our case. The calibration results for the SI model reflects similar results for the weights associated with its factors as follows.

For calibration results, the aggregated calibration for all household types showed better results than the disaggregated calibration. This is undoubtedly because the disaggregation process dissected the sample into small groups of calibration samples that would be unrealistically biased. While, the correlation coefficients for the disaggregated calibration for both training data (2010) and prediction data (2015) were at (r= 0.660) and (r= 0.698) respectively. The aggregated calibration for the same data were almost identical at (r= 0.720) and (r= 0.723). These results do not only show the marginal superiority of the aggregated calibration but it also indicates a reasonable stability in the strength of prediction over a short time horizon. Given that the aggregate calibration is better and that it uses a single weight for each factor, it will be further discussed in this section. The associated weights for the count of issued residential building licenses ($\boldsymbol{x_l}$), the average square meter price for residential land ($\boldsymbol{x_p}$) and the count of unique commercial activities ($\boldsymbol{x_s}$) were weighted at 4.157, 0.294 and 0.357. The weight of residential mobility is substantially higher than all the three other variable which, agrees with the strong correlation of mobility trends and issued residential licenses shown earlier. For the remaining factors of distance weight ($\boldsymbol{x_d}$), exponential distance fraction ($\boldsymbol{d_{ij}^{\beta}}$), scale $\boldsymbol{a^t}$ and threshold the rates were at 0.076, 0.709, 0.952 and 0.205 respectively. See Table 39.

Table 39 the calibration weights for the aggregated household type in the SI model.

Factor name	Factor symbol	weight
Residential building license	x_l	4.157
Residential land prices	x_p	0.294
Commercial activity licences	x_s	0.357
Distance	x_d	0.076
Distance $\boldsymbol{\beta}$	d_{ij}^{β}	0.709
Threshold trigger		0.952
Scaling factor	a^t	0.205

It also has been noted that distance weights x_d and d_{ij}^{β} produce a smooth increase in distance sensitivity that will eventually get damped more like the properties of a logarithmic curve. See Figure 36 for testing the outcomes of SI distance weights for an illustration of the produced curve.

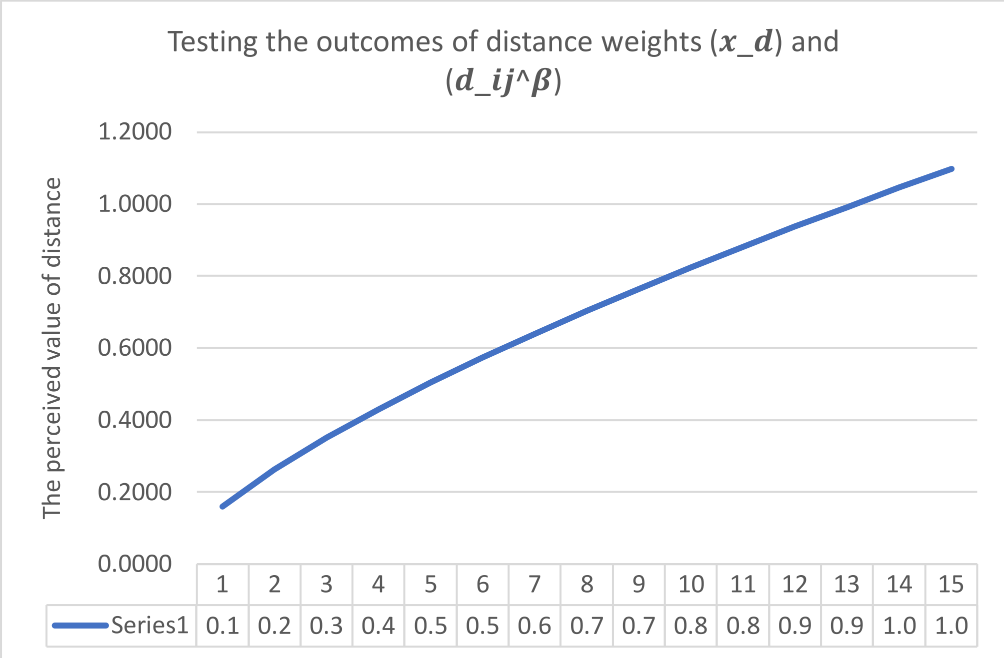


Figure 36 testing the outcomes of SI distance weights.

Last but not least, the threshold value at 0.952 seems to indicate a willingness by households to consider districts with a slight decrease in benefit but not below the current level anyway as the threshold can be rounded to 1.

A.2 The sub-model of NN modelling (location)

The NN model is the second approach to model residential mobility. A Household's attributes, including its current location, are used as inputs for prediction. However, this approach does not consider the changes at the district level. Therefore, this model cannot handle the establishment of new districts, which makes this model unsuitable for long-term horizon forecasting in which the number of districts is expected to grow. This sub model consists of two separate parts to estimate the probability of moving to all other districts and to estimate the living duration of households given their status. Like the SI model, the calibration process considers disaggregated household types by age, number of children and current location. In the case of absence of a household type in the sample, an aggregated household type is used to cover the absence. Also, due to the expected increase in modelling time due to calling external functions for many agents, a premutation approach was adopted to insert all the NN model results into AnyLogic as a hash map list. This approach was successful in querying the results for each case with reasonable retrieval speed for thousands of agents. The query of results is similar to the

SI approach, where all types of households were uniquely coded by concatenating the attribute values in a string. The location model runs as follows:

First, check if the household type is within the calibrated with disaggregated household by age, number of children and location or just age and number of children. Then, query the associated probabilities for moving to all districts from the dedicated HashMap. After that, copy the probabilities to the districts' probability list.

A.3 The sub-model of NN modelling (Duration)

For the duration mode, the same steps are applied to query the probabilities from the HashMap lists based on the household's type. However, instead of returning a list of probabilities, the duration values are expected.

For NN, both location and duration models were created based on supervised learning for a multi label classification problem. These models utilise both Keras and TensorFlow libraries to create a Sequential model. The proposed models consist of 4 layers as follows. First, the input layer with nodes equivalent to the number of input variables. Moreover, Rectified Linear Unit (ReLU) is used as for activation. Second, two additional hidden layers with the same number of input variables for nodes and ReLU activation functions. Third, the output layer uses a SoftMax activation function to produce the probabilities for each expected output. The model is compiled with the categorical cross entropy loss function. Moreover, the gradient descent algorithm "adam" is used as an optimizer function. Furthermore, the classification accuracy is calculated using "top k categorical accuracy" function[4] with the value of k being tested for top 1, 2, 3, 4 and 5 probabilities separately.

The differences between the location and duration models lies in the calibration process. While the duration model is assumed to predict the duration of living in a home based on a life cycle of households. This is expected mainly to depend on the household's attributes

---

[4] the "Top k categorical accuracy" for the duration model was modified. This is because, it has been found that zero probabilities would be included despite being hypothetically excluded and that duplicated probabilities are included with a fixed order. In other words, in the case of three probabilities with each being 33.3%, the top 1 categorical accuracy will always pick the last option despite the equal probability which should give a uniform distribution of picking one of the options. The modifications to solve these issues includes removing choices with zero values and randomly changing the duplicated values to get the effect of uniform consideration of options.

as younger households are expected to be more prone to change their homes with the changes in the number of children or income and verses versa. The location model, on the other hand, is more dependent on the trends of residential developments which is barely related to the household life cycle. Therefore, the sample split is handled differently in these models. In the duration mode, the entire sample is considered with no year-based split. Moreover, the calibration is based on a 50% shuffled data split. Then both epochs and batches were set to 200. However, for the location model, as the trend shifts with time, only the last 10 years were considered from the sample observations. Then, the data was further split as training data covering the years (2006 – 2010) and prediction years covering (2011 – 2015). Moreover, both epochs and batch size were set to 5000 and 200 consecutively.

The predictability goodness for the NN model was tested at the level of fitting the NN model and the overall simulation outcomes. Evaluating the simulation outcomes will enable the comparison of the prediction goodness among the different models available for the user.

First, fitting the duration model was done by splitting the data by 50% between training and testing, and the first prediction option (top k=1) had a 45% chance of being true. This probability of true prediction increases with the order of top k categories to reach about 90% if we consider the first 5 options (top k=5). Second, the location model data was split according to the years of movement which is broadly categorised as 2010 training data and 2015 testing data. For the training data, the first option (top k =1) had a probability of correct prediction close to 55% and, up to 93% when the first 5 options (top k=5) are considered. It is worth mentioning that the first 5 options are chosen from 150 different alternatives. This aspect shows a strong ability in narrowing the space of alternatives by around 97%. However, the testing data showed less remarkable results. For the first option (top k = 1) the predictability plunged to less than 10%. However, with considering the second option the probability of true prediction increases to about 47% and, improves to reach above 70% when considering the first five options (top k=5). The sharp decrease that occurs with the first option (top k= 1) is caused by the inability of NN models to predict the moves based on changing derivatives. In other words, the NN model is trained to predict fixed outcomes based on a temporal trend and if the trend and its outcomes change, the NN model will not change automatically (i.e.it has no capacity to sense and react accordingly). Therefore, the outcomes will continue to be same despite any changes. In our case, it could be said that the most attractive districts in year 2010 have changed in year 2015. Consequently, the usual first options (top k=1) in year 2010 do not hold to be true for 2015 which cause the drop in predictability. Moreover, because the time horizon is relatively short (around 10 years) between the testing and prediction data, it can be seen that the model predictability is above 70% for the first 5 options (top k= 5). See Figure 37 for the actual trends of residential mobility destination locations for 2010 and 2015.

On another front, to evaluate the actual application of the NN location model to estimate residential mobilities as part of the ABM framework, the Pearson Correlation Coefficient was used to test the correlation between the actual and predicted inflows at the districts level for the years 2010 and 2015. The testing was done in two ways: moving all households with the location model only and moving households based on duration and location models combined. The results for moving all households unconditionally by the duration model for 2010 and 2015 scored (r= 0.68) and (r= 0.76) respectively. The conditioned mobility by duration for the same periods scored (r= 0.64) and (r= 0.73) respectively. The decrease in r values for conditioned mobilities are expected as the sample exclusively contains moving households, Also, the fact that the decrease is marginal indicated good predictability of moving households. However, the surprising result comes from the fact that the correlation coefficient for year 2010 is somehow lower than year 2015, despite the higher predictability of year 2010 training data over the testing data of year 2015. Nevertheless, it can be argued that these measures are not directly comparable. While the Pearson coefficient is largely an indicator of how close the predicted inflows associate with the actual inflows, the top k evaluation metric highlights the probability of finding the correct answer within a limited range of options. Also, in this testing, the final mobility destinations are based on weighted probabilities and then evaluated for their correlation to the actual flows. Therefore, the two tests should be considered as complementary more than contradictory for evaluating the models. The author believes that different evaluation metrics should be examined including the use of visual comparisons of outcomes to better understand the models' outcomes. See Table 40 for evaluation results for both duration and location NN models using Top k category and Pearson's test. Also, Figure 37 for the visual comparison of actual and predicted mobilities using both NN (DL) and SI models for the years 2010 and 2015.

Table 40 evaluation results for both duration and location NN models using Top k category and Pearson’s test.

Model	**Predictability of Duration model (k=10)**	**Predictability of Location model (k =150)**		**Pearson’s R-value for simulated district inflows (Unconditioned by duration)**	
Evaluation metrics	Testing 50% split	Training 2010	Testing 2015	2010	2015
Top k 1	0.45	0.54	0.073		
Top k 2	0.66	0.75	0.47		
Top k 3	0.78	0.85	0.62	0.68	0.76
Top k 4	0.85	0.90	0.69		
Top k 5	0.91	0.93	0.73		

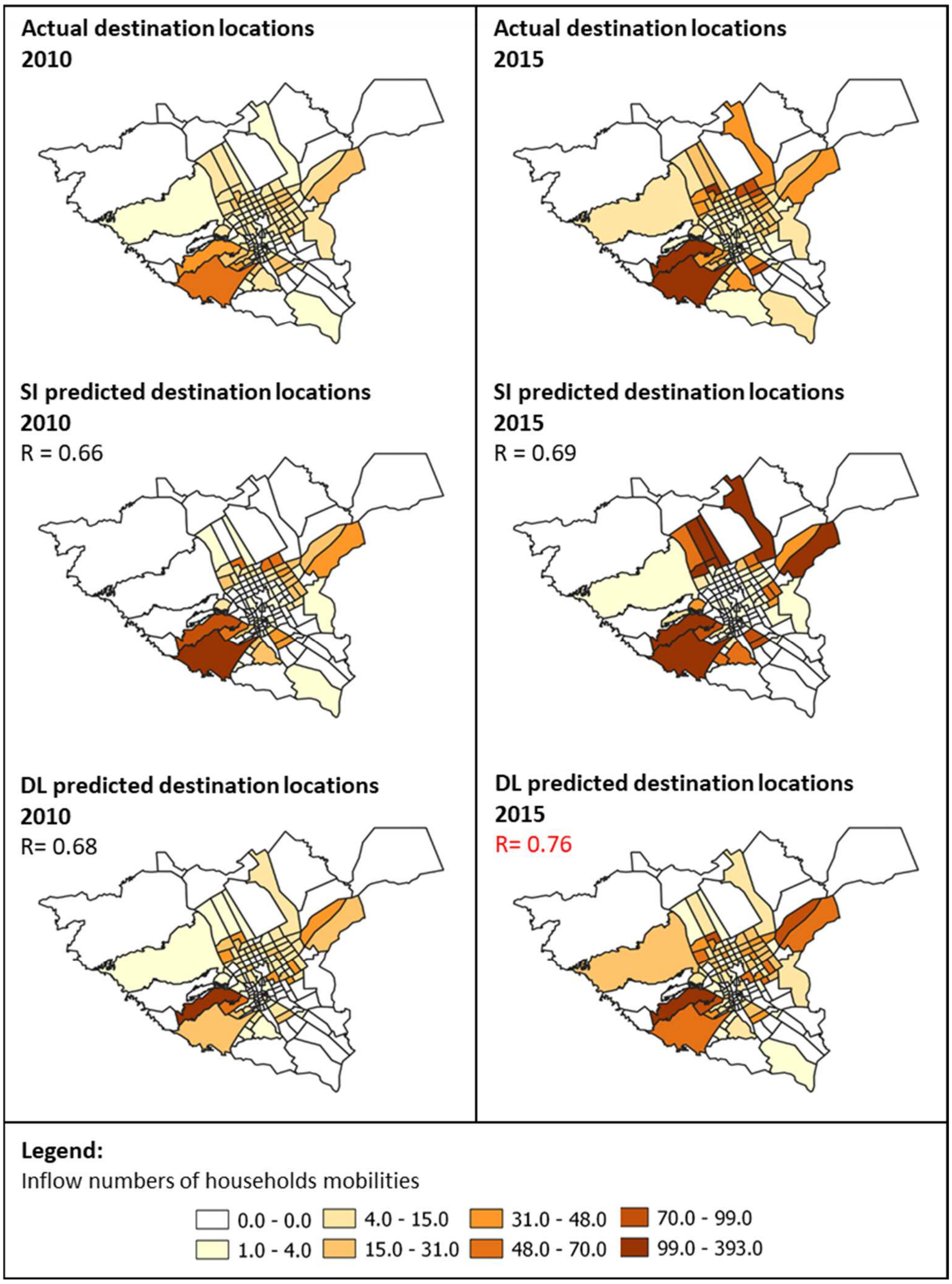


Figure 37 The visual comparison of actual and predicted mobilities using both NN (DL) and SI models for the years 2010 and 2015.

Figure 37 shows the actual destination locations of households' mobilities in the years 2010 and 2015 and then the predicted location of households' mobilities using NN and SI models. In year 2010 both NN and SI models replicated the actual mobilities fairly well with approximately similar R values. The NN model has the advantage of not emptying the central part of the city. Moreover, for year 2015 it can be visually noticed that the SI mobility predictions were strongly oriented to the outskirts of the city with a central emptying trend where the NN predicted mobilities have a more balanced distribution. At this point, it can be said that many of the inner small districts in the NN model were closer to the actual mobilities therefore, the regression-based comparison was in favour of the NN model. Although the SI model was not very far from reality, it is believed that it could achieve better results with calibration refinement.

B. The sub model of marrying for the first time

The marriage for the first-time function is coded within the male children class of agents. This function is built upon the assumption that never married adults are most likely to marry a never married before partner. This function has three conditions to proceed: the age of agent is above 15, the randomly weighted probability by age is true and the agent is active. If these conditions are met, then the agent is ready to get married and, the next stage is to find a partner. Finding a female partner from the female children class of agents is conditioned by the age being 15 and above, while the maximum age difference is not more than10 years, the randomly weighted probability is true and the agent as active. Then, if an agent is found, a new household agent is established and both agents will be cast as Father and Mother agents to be connected partners in the newly established household.

C. The sub model of marrying for the second time

For each willing unmarried male-headed household, with likelihood probability to marry widow, select a random existing single female-headed household for which the age of the woman is equal or younger than the proposing man. If such a household is found, the woman is connected to the male's household in the mother role. All of the female-headed household members are removed from their existing household and added to the male-headed household; they are moved to the male-headed household's location and the now-empty household are removed.

If, on the other hand, the householder chooses not to marry a widow/er, then a never married before partner will be looked for with the same age conditions. If a suitable partner agent is found then, the agent will be disconnected from its current household. After that, the partner agent will be cast as an adult agent of type Father or Mother based on its gender to be linked to its new household. In this case, partner's original agent is removed and considered inactive, while the newly created agent by casting will join suitor's home location.

There is a possibility that when a sibling is promoted as a household head, he will get married using the remarrying function instead of the marrying for the first time. This only affects that agent by prioritising his marriage to a previously married partner (most likely a partner with children) than marrying a never married before partner. No testing of this matter effect on the population size was done as it is expected to be marginal, given the low chances of this to occur in the first place. This issue could be eliminated in future work by adding more controlling parameters.

D. The sub model of Birth

The function of birth is embedded within the agents of class mother only. Moreover, this function is called for all active mothers between 15 and 50 years old who are connected to a husband agent through a household container. If these conditions are met, two separate birth rates for both female and male fertility rates are checked. These rates are based on the mother’s age and the birth’s gender. Based on the birth probability. a new baby agent may be created with age 0. The new child agent will be connected to their mother’s household agent. Finally, the baby agent will be located in its household district.

E. The sub model of Death

All individual agents have this function implemented within them. This function does not remove the agents completely from the modelling environment. However, it changes the status of dead agents to be inactive. This approach was preferable to ensure that the unique identification numbers of all agents remain the same over the simulation time horizon which enables better tracing and debugging. Therefore, the condition for checking active agents became essential in calling any function in the model. For agents to be removed, their death function must return a true value based on age weighted probability. Then, any dead agent is disconnected from all its links to other agents (i.e. children, parents and households) and released from its location to the cemetery as its new location. The collective agents of type household do not have a death function to be inactive. However, when the household agent becomes disconnected from any individual agent, its is then removed to the cemetery location.

F. The sub model of Ageing

All individual agents have an age parameter called age. This parameter is increased by 5 years for each step of the model if the agent's status is active.

G. The sub model of Divorce

The divorce function is part of the household agent. This function is called only when both parents are connected to the household. Then, two parameters are checked: First the divorce rate and then, the mom custody rate while the divorce rate is a function of the household head age. With probability 0.7, mothers will retain control of the Household and the children, and a new Household will be created for the father to join. Otherwise, the mother will create and move to a new Household while the father and children stay in the original Household. This process includes the following steps: create a new household agent, disconnect the husband agent from the current household container and connect it to its new household agent, update the characteristics of new household agents head with the fathers' characteristics, call the residential mobility function to determine the new location for the father agent and release the agent from its current location to the new location. On the other hand, if the mom custody rate is false then the mother agent will be linked to a newly created household agent whereas children will remain with their father. The same steps are executed but for the wife agent.

H. The sub model of children custody

This function is coded within agents of type household and activated when a household agent lacks both Father and Mother roles, which could occur in the event of the parent death in a single parent household or, more improbably, both parents dying during the same step. To deal with the issue of no parents in a household, a deterministic approach is followed with five consequent cases. First, a randomly chosen independent male sibling (i.e. married brother) will take all his siblings under his custody. Second, if a male sibling is not available then, a randomly chosen independent female sibling (i.e. married sister) should take the custody role. Third, alternatively, if no independent siblings are available, then an adult male sibling from the dependent children should be randomly selected, but conditioned by his age to be above 25 (i.e. expected to be working) to be assigned as the household head. Fourth, if the male adult siblings do not exist, then an adult female sibling should be selected as the household head with the same age criterion. Fifth, if no adult siblings are available to take responsibility, then the function will search for a random household conditioned by head age to be above 25 to take the custody role, based on the assumption that a relative or fostering family will fulfil this obligation. The process associated with the aforementioned cases is as follows:

I. The sub model of deciding the household head

This function is part of the household functions. Due to the nature of continuously changing household attributes due to events like marriage, divorce or death, the status of the household head changes accordingly. Therefore a function is needed to check and update the status of the household head and also, to account for the possibility of parental partnership in decision making. This function comes at the end of all population changes which determine the household head of each family. In the cases of a single parent household, the function will simply ensure that the household head characteristics are updated with the single parent attributes. However, when both parents are connected to a household, the qualities of either the father or the mother, or both partners together, are taken into account based on a weighted probability. The consideration of partnership is currently done by making the male characteristics as dominant for the household head except for the income parameter which equals the sum of both parents' income. The partnership relation was considered to highlight the opportunity to include such an aspect in dealing with households' data. However, it was not further refined by adding another parameter to indicate this feature in households. This is because both partnership and income were not considered in the training data of residential mobility.

J. The sub model of adding migrated population

This function was included as an exclusive event outside all individual and collective entities. Because migrated population are added and based on an expected number, the function must get the expected number of migrated populations in the first place. In our case, the number of migrated populations is assumed to continue at the current levels. However, we add some randomness by (+/- 10%) to the fixed number of migrants to mitigate the rigidness of this assumption. After getting the final number of migrants, an equivalent number of new households will be added to the population in an iterative manner. For each newly created household two main characteristics must be defined which are age and gender. Currently, the available data provides information about annually migrated individuals without further information about the distributions of age or gender. However, to account for the distribution of these aspects, both age and gender are chosen based on weighted random probabilities. Then, the first location of this newly created household agent is decided based on the active residential mobility sub model as explained in the sub model A earlier in this section. After defining the main characteristics of the migrated household and its initial location, the individual agent associated with this household is created according to the defined age and gender. Nevertheless all the other characteristic of the new household are left empty as they do not play a role currently in agent behaviour. Then the household agent is released to its initial location followed by its individual agent.

These eight sub models represent the core of the ABM model which works in harmony to simulate the aspects of population growth, social change and residential mobilities. The model has a generic framework that was developed intentionally to simulate population change at the district level of any city and population beyond the scope of this research. Also, it is necessary to be largely flexible to include or exclude key variables and as income or even social characteristics such as polygamy when needed.

4.2.3 Optimisation at district level

Unlike the city-level optimisation, which uses the area per student mainly as a provisional standard, district-level optimisation could be implemented in two distinct ways. The first optimisation strategy is identical to the city-level optimisation, which optimises the area per student standard. The second optimisation strategy attempts to explore an appropriate combination of pre-designed school prototypes that would fit on the current land parcels in order to accommodate projected future student surpluses. The latter strategy is comprised of two functional sub-strategies that assume the number of lands to be either limited (v1) or unlimited (v2). However, only the second main strategy will be studied at the district level, as influencing the MOE’s selection of a building prototype is more likely than the modification of ADA's current planning criteria. Moreover, to demonstrate the two optimisation sub-strategies without adding further district-level complexity, if the outputs of one strategy are preferred more than those of the others, it is possible to replace the optimisation functions. Before explaining this approach in detail, few points must be considered:

1) The education system in Saudi separates the pupils based on gender.
2) Some land parcels are large enough to accommodate more than one building.

3) The number of land parcels varies among districts and exceeds 40 parcels in some large non-subdivided and underdeveloped districts.
4) The MOE has developed 22 prototypes for building new schools. These prototypes are classified by the number of education stages considered for each prototype, the building footprint area and the student capacity by stage. See Table 35 for the classification of prototypes.

Consequently, the main notion for optimising the provision standards for schools which was introduced earlier in Equation 19 must be adapted differently to find a combination of parcels' specific prototypes that would accommodate the surplus of students over time proactively. Therefore, the optimisation criterion will can be written as:

$$\sum^{as} count_{t+n}^{x \in p^*} * capacity_{t+n}^{x \in P^*} > surplus\ students_{t+n} \tag{33}$$

Equation 34 annotation of filtering the school building prototypes using land area.

$$P^* = P \cap (PL \in \ddot{L}) * MBFAR \tag{34}$$

Equation 35 the calculation of net land area using districts' specific parcels area.

$$net\ available\ land = \sum (PL \in \ddot{L})\ * s \tag{35}$$

where:

a	land area dedicated to education purposes.
s	the land preservation factor (precautionary set to 0.3)
$count_{t+n}^{x \in P^*}$	the count of prototypes of type x from the prototypes list P^* in time $t+n$
$capacity_{t+n}^{x \in P^*}$	the student's capacity of prototypes of type x from the prototypes list P^* in time $t+n$
P	a list of all unique possible combination of schools' prototypes
P^*	a land specific prototypes' list that is filtered by the maximum building footprint area $MBFAR$ (subset of P which is constrained by existing available land)
$\ddot{L}$	a list of all education allocated land parcels in the targeted district
$PL \in \ddot{L}$	the parcel of land PL from the list of parcels $\ddot{L}$
$MBFAR$	The factor of maximum building footprint.

Equation 33 shows that the optimisation function is a product of summing the capacity of all parcels' specific building prototypes, which is limited by the minimum value of surplus students and the maximum value of net land area for each district. Moreover, the list of all unique possible combination of school prototypes, which is annotated as P, is created by expanding the original list of prototypes in Table 35 to a unique list of prototypes combinations that are non-duplicated by stage. For example, if a school prototype is designed to accommodate elementary and middle stages, only the secondary stage prototypes will be used to create new unique combinations. The number of unique combinations in list P is 184 combinations with footprint areas ranging between 1002 to 10944 Sqm. Also, the number of students in all stages range between 630 to 2520 students. See Table 41 for a random sample of combined school building prototypes with its unique ID, total footprint area and total capacity.

Table 41 A random sample of combined school building prototypes with its unique ID, total footprint area and total capacity.

unique ID	Total area of combined prototypes (Sqm)	Total capacity of combined prototypes (student)	Elementary stage capacity (student)	Middle stage capacity (student)	Secondary stage capacity (student)
36	1002.64	630	630	0	0
71	1796.86	630	630	0	0
1	1002.64	840	840	0	0
73	2799.5	1260	630	630	0
184	5472	900	0	0	900
178	1702	660	220	220	220
42	2005.28	1170	630	0	540
77	2799.5	1170	630	0	540
153	10944	1890	495	495	900

Table 41 shows that the combinations of prototypes does not have a strong linear correlation between the total capacity and footprint area at r= 0.55. Moreover, the education stages could include any combination of the three education stages without considering their ageing order. Unfortunately, these features add more complexity for the optimisation process in finding the optimal solution. The optimisation function is applied using the following sequence of programmed events:

1) At the warming up stage:
    a) The targeted district for optimisation will be manually defined.
    b) The list of unique prototype combinations will be manually filtered for each land parcel within the targeted districts using its area.

c) The population change model will run until the targeted year. This model will run under fixed random seeds.
d) All data related to school age population are stored in separate variables by education stage.

2) For the optimisation parametric variation
    a) For each land parcel, vary the Boolean gender parameter between male and female.
    b) For each land parcel, vary the discrete parameter of possible prototypes combinations. The number of possibilities will be based on the range defined by area in step (1.b).
3) For the optimisation objective, minimise the number of surplus students.
4) At the end of each optimisation iteration, check the set of conditions to accept or reject the solution.
5) Repeat the previous steps for a defined number of iterations.

The optimisation process notably requires some manual inputs, this is because:

1) Each district is optimised independently.
2) The number of created parameters for gender and prototypes are associated with the number of parcels within each district.
3) The number of unique prototypes that can be built in each land parcel is conditioned by the land area. Therefore, two solutions could be applied. First, filtering the possible prototypes for each land parcel by its area. Second, we add a condition to check that the prototype footprint area is less than the land maximum built-up footprint area. The first approach is believed to reduce the optimisation time by narrowing the solution space.
4) For the land preservation factor, the author believes that such a factor at the parcel level must be only considered when there are more than two parcels. This is because we optimise each parcel for each of both genders. Moreover, subtracting the land preservation factor must be calculated at an aggregate level rather that the parcel level to make use of the maximum building footprint in each parcel.

These aspects at the current time can mainly be adjusted manually for the optimisation engine process. Therefore, the optimisation process tends to be manually controlled. Overall, this optimisation approach is expected to perform well for a small number of parcels such as three or below, but when the number of parcels increases, the number of possibilities increases exponentially. For example, if the number of permutations for 2 land parcels and 184 prototypes to choose from is 33672, then, the number of permutations for 4 land parcels will be 1109223024. Moreover, these numbers would double if we considered both genders for each land parcel. Thus, the optimisation process for a larger number of land parcels can be viewed as an attempt to find the good solution rather than a process guaranteed to produce the very best solution.

4.2.3.1 Optimisation objectives and variables

Two optimisation functions were created for the outcomes at the district level. This is because land-specific optimisations necessitate declaring two optimisation parameters for each specific land, resulting in a total of 118 parameters in one of Riyadh's city districts, given the highest number of 59 education-related lands. In addition, to ensure a land-specific optimisation process, a number equal to the number of lands must be specified as area constraints. The issue of land-specific parametrisation increases the size of the optimisation solution space, which has an impact on the optimisation performance. As a result, a more liberal optimisation function was developed as an alternative to the land-specific function with a fixed number of 44 parameters. The second function converts the demand for surplus students into a supply of school prototypes without taking into account land-specific constraints. This function can be used when the demand for schools surpasses the first optimisation function's predefined limit. To further clarify these two functions, consider the following:

4.2.3.2 The land specific optimisation function

The initial optimisation function focuses on a solution that is specific to the land. It has the ability to optimise up to ten different land parcels, either actual or hypothetical. In this approach, each land is assigned two parameters: one for selecting a school prototype by ID and the other for determining the gender of the targeted pupils. To determine the best solution, the optimisation algorithm will experiment with 184 different combinations of school prototypes for up to 10 land parcels. Although there are 22 prototypes, they were synthesised into 184 non-overlapping education stages to account for the possibility of more than one prototype being established in each land parcel. This would make better use of any available land parcel. However, in the case of huge land parcels exceeding the combined prototypes' maximum area, a case-specific school design may be required. Furthermore, a land-specific restriction is introduced to make sure that the prototype footprint fits the exact land area, allowing this function to target actual land parcels. Otherwise, the land area will be unrestricted, allowing the optimisation process to suggest the required building prototypes, which may then be translated into the required land area to be supplied or acquired.

To deal with the temporal dimensions of changing demand over the simulation time horizon, a basic smoothing step is applied to the number of students in each iteration. The smoothing function applied is a Cumulative Moving Average which works as a fluctuation limiter and brings the total average to a closer point with respect to the general trend direction.

Objective function:

$$\min \sum_{{}^{\circ}k=1}^{10} abs[(X + C) - Y_{{}^{\circ}K}] \tag{36}$$

$$X = \begin{bmatrix} x^{m,ele} \\ x^{m,mid} \\ x^{m,sec} \\ x^{f,ele} \\ x^{f,mid} \\ x^{f,sec} \end{bmatrix} \tag{37}$$

$$\underset{\substack{s.t.\ p_i^* \in [0,184], \\ i \in [1,10],\ \ 10 \geq u \geq 1, \\ g \in [m,f],\ \ g = m \vee f}}{C^*} = \begin{bmatrix} p^{*g,ele}_{\ i} & p^{*g,mid}_{\ i} & p^{*g,sec}_{\ i} \\ \dots & \dots & \dots \\ p^{*g,ele}_{\ u} & p^{*g,mid}_{\ u} & p^{*g,sec}_{\ u} \end{bmatrix} \tag{38}$$

$$\underset{\substack{i \in [1,10],\ \ 10 \geq u \geq 1}}{C_{s.t.\ \left(p_i^{*m} \vee p_i^{*f}\right) = \nexists = 0}} = \begin{bmatrix} c^{m,ele} = \sum_{i=1}^{u} p^{*m,ele}_{\ i} \\ \sum_{i=1}^{u} p^{*m,mid}_{\ i} \\ \sum_{i=1}^{u} p^{*m,sec}_{\ i} \\ \sum_{i=1}^{u} p^{*f,ele}_{\ i} \\ \sum_{i=1}^{u} p^{*f,mid}_{\ i} \\ \sum_{i=1}^{u} p^{*f,sec}_{\ i} \end{bmatrix} = \begin{bmatrix} c^{m,ele} \\ c^{m,mid} \\ c^{m,sec} \\ c^{f,ele} \\ c^{f,mid} \\ c^{f,sec} \end{bmatrix} \tag{39}$$

$$Y_{{}^{\circ}k\ s.t.\ {}^{\circ}k \in [1,10],\ \ 10 \geq n \leq 1} = \begin{bmatrix} y^{m,ele} = \frac{y^{m,ele}_{{}^{\circ}k=1} + \dots + y^{m,ele}_{{}^{\circ}k=n}}{n} \\ \frac{y^{m,mid}_{{}^{\circ}k=1} + \dots + y^{m,mid}_{{}^{\circ}k=n}}{n} \\ \frac{y^{m,sec}_{{}^{\circ}k=1} + \dots + y^{m,sec}_{{}^{\circ}k=n}}{n} \\ \frac{y^{f,ele}_{{}^{\circ}k=1} + \dots + y^{f,ele}_{{}^{\circ}k=n}}{n} \\ \frac{y^{f,mid}_{{}^{\circ}k=1} + \dots + y^{f,mid}_{{}^{\circ}k=n}}{n} \\ \frac{y^{f,sec}_{{}^{\circ}k=1} + \dots + y^{f,sec}_{{}^{\circ}k=n}}{n} \end{bmatrix} = \begin{bmatrix} y^{m,ele} \\ y^{m,mid} \\ y^{m,sec} \\ y^{f,ele} \\ y^{f,mid} \\ y^{f,sec} \end{bmatrix} \tag{40}$$

Where:

X	is a given list of all current schooling capacities by gender [m , f] and education stages [ele, mid, sec].
C I	is a calculated list of all newly add capacities by gender [m , f] and education stages [ele, mid, sec]. The C list is calculated based on the C^*.
C^*	is a calculated matrix that stores all the capacities generated by the optimisation alternatives $p^{*g,}_{i}$ for each land parcel that is annotated by the subscript i. Each row in the C^* matrix represents the capacity outcomes of a chosen prototype p^* specific to a land parcel i by gender [$m \vee f$] and education stages [ele, mid and sec].
$p_i^{*g,s}$	is the additional school capacity added from the building prototype p_i^* given the gender g = [$m \vee f$] and the associated education stages [ele, mid and sec]. p_{ij}^* is chosen from a predefined list of 184 options for each parcel of land.
$Y_{\circ K}$	is a calculated list of cumulative moving averages of all students counts by gender [m , f] and education stages [ele, mid and sec].
g	is gender which can be either male (m) or female (f)
s	is the education stage which can be (elementary = ele), (middle = mid) and (secondary = sec).
i	the targeted land
$\circ k$	The iteration's order.

4.2.3.3 The non-land specific optimisation function

The second function disregards any specific land limits by accounting for 44 distinct alternatives that reflect all possible schooling prototypes. There are 22 prototypes in total, each of which is duplicated to cover both genders separately. Nonetheless, the optimisation could be constrained by prototypes total area or total number of schools. However, the optimisation itself does not have a bias toward supply reduction in general. The mathematical difference lies in changing the C^* parameters in the first optimisation function (v1) from the utilisation of ten land-specific schools buildings' alternatives to the required number of schools from the 22 building prototype alternatives. This solution was applied as follow:

$$\underset{\substack{s.t.\ p_{ij}^* \in [0,22], \\ i \in [1,22],\ \ 22 \geq u \geq 1, \\ j \in [1,6],\ \ 6 \geq n \geq 1, \\ m \in [1,\infty],\ \ \infty \geq m \geq 0, \\ f \in [1,\infty],\ \ \infty \geq m \geq 0}}{C^*} = \begin{bmatrix} m_i\, p^{*ele}_{ij} & m_i\, p^{*mid}_{ij} & m_i\, p^{*sec}_{ij} & f_i\, p^{*ele}_{ij} & f_i\, p^{*mid}_{ij} & f_i\, p^{*sec}_{ij} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ m_i\, p^{*ele}_{uj} & m_i\, p^{*mid}_{uj} & m_i\, p^{*sec}_{uj} & f_i\, p^{*f,ele}_{uj} & f_i\, p^{*f,mid}_{uj} & f_i\, p^{*f,sec}_{un} \end{bmatrix} \quad (41)$$

Where:

C^*	is a calculated matrix that stores all the capacities generated by the optimisation alternatives $m_i\, p^*_{\ i}$ and $f_i\, p^*_{\ i}$ for each school prototype of type P_i^*. The numbers of each prototype to be provided by male and female students is distinguished by the parameters m_i and f_i respectively.
ele	is elementary stage
mid	is middle stage

sec is secondary stage
i the unique identifier for the 22 building prototypes.
j the unique identifier for the education stages ele, mid and sec.

Equation 41 heavily modifies Equating 38 to introduce the prototype specific alternatives with the addition of a multiplication parameter for each prototype by gender to estimate the numbers requited of each of the 22 prototypes by gender. The capacity of each education stage is a product of multiplying the multiplication parameters m_i and f_i by the unique prototype P_i^*. The remaining steps in Equations 39 and 40 are the same.

4.3 The differences between city and district level optimisation

It can be said that the differences among the three optimisation approaches introduced in this study are significant. At the city level the typical area per student standard is optimised with a strong focus on the amount of consumed land at multiple 5 year time steps. At the district level, a new approach that utilises the predesigned building prototypes and the specific land parcels available is applied for a defined time horizon. Moreover, the perceived advantages of using the city level could be in its simplicity to estimate the total amount of needed land area and the enhancement of strategic planning by considering 5 year time steps for the provision of services. However, this approach is not sensitive to the distribution of population and the practical side of predesigned facilities and available lands. On the other side, the district level optimisation functions does cover the shortcomings of the city level approach but, limits the power of strategic plans as it average the supply and demand for a single and usually prolonged time horizon.

The following chapter will present the outcomes of the various methods discussed in this section at both the city and district levels and compare some of their outcomes in order to illustrate their degree of resemblance.

Chapter 5

Results

This section consists of two main subsections which are the city level and the district level. Within each subsection the experimental design, optimisation outcomes and the results are explained. The city level will be demonstrated first then, followed by the district level.

5.1 City level

The model outcomes at the city level can be viewed as the most aggregated form for optimising the provision standards of schools. Moreover, this approach can be considered as the benchmark for population projection given its use of conventional CCM. The user experiments were designed to investigate different inputs for the base year population counts, fertility rates and migration rates. The different scenarios will be discussed to highlight their expected outcomes and the model sensitivity to its variables. However, the focus will be on the three most likely scenarios to occur.

5.1.1 Optimisation inputs

At this stage it is important to justify the existence of two base year datasets as a choice for input data. The original population estimates by ADA for Riyadh city is the only available data that is disaggregated by district, age and gender. However, for the purpose of using this data in an ABM environment it could be either synthesised as household-based population or used in its raw form as a sampled household-based population. Although, population synthesis is a commonly used approach in urban simulation, the use of sample population seemed promising as they consist of ready to use household-based population data and would reduce the demand for computational power. However, the issue of inconsistency in rates to convert raw sampling data into population estimates that are matching the original ADA estimates is hard to overlook. Therefore, the raw sampling data had to be adjusted as described in section C.3 The data curation model. The reweighting process was able to resemble the original estimates to a great extent but, it does not achieve perfect matching of population counts. Therefore, to create comparable outcomes from the two models developed in this research (i.e. the models of city level and the district level), the adjusted population estimates should be used as an input for the city model to be comparable with the adjusted sampling data that is used as an input for the district model. Consequently, the difference between both original and adjusted population estimates must be understood in terms of initial and projected values.

This section is divided into three subsections as Riyadh city population, targeted schools age population and the optimisation of standards.

5.1.2 Population growth:

For the population growth, both the original and adjusted ADA population estimates were initially considered as base year population. Given the used approach for adjusting the ADA sampling population, some difference is expected between the original and adjusted data sets. The adjusted population has a 2.16 % increase in total population counts. However, the age specific differences for both genders averages 3.41% with a standard deviation of 6.39%. Also, the gender specific differences are more significant in deviation, while males average 1.67% with a standard deviation of 14.47 %. Females averages 2.93% with a standard deviation of 9.41%. See Table 42 for comparison between original and adjusted ADA 2016 population counts for Riyadh city.

Table 42 shows that there are some significant variations between the original and adjusted population estimates. The variation level in some cases is drastically high around a 25% increase or decrease. However, to explain the expected effect of these variations, male then female population will be discussed. On the one hand, male school age population which is between age 0 and 19 years, tend to be deflated by an average of 8.64% in the adjusted population. Accordingly, the first three iterations of population projections might have less demand of land area by male students. However, given that birth counts are based on female counts, the magnitude of variation in males beyond age 19 will not affect the targeted school's population in coming years. It would be necessary to close the variation gap if other older age groups are targeted by the model but, this is not the case for this research. On the other hand, female schools age population tend to be inflated by 3.67% on average to cause more demand for education land area. This matter can be seen to have some sort of cancelling effect to the deflation in male school age population if the gender factor is excluded/overlooked. However, given that both counts of current women in childbearing years and women in schools age population, which will become part of the reproduction population in the coming years, are by and large inflated by 7.27% on average. The numbers of school age population for the adjusted population are anticipated to be higher in future projections. Consequently, the standards optimisation based on the adjusted population could increase standards capacity by 7% over the base line of original population estimates. Therefore, it is believed that the original data should be used mainly to examine optimisation results under various factors. Then testing the outcomes of the most likely scenarios on the adjusted population to highlight the necessity for re-optimisation under slightly different population structures with the anticipated 7% increase in needed capacity.

Table 42 comparison between original and adjusted ADA 2016 population counts for Riyadh city.

Age group	Original base year		Adjusted base year		Percentage difference		
	Male	**Female**	**Male**	**Female**	**Male**	**Female**	**Total**
00-04	176515	149481	155200	149500	-14%	0%	-7%
05-09	233231	196153	214100	204300	-9%	4%	-3%
10-14	270374	227639	263600	245100	-3%	7%	2%
15-19	289171	245939	315900	270400	8%	9%	9%
20-24	284638	242363	301000	257000	5%	6%	6%
25-29	212246	179726	208200	184300	-2%	2%	0%
30-34	149477	127097	137900	126300	-8%	-1%	-5%
35-39	137938	117200	109700	125700	-26%	7%	-8%
40-44	106439	91370	92300	109200	-15%	16%	2%
45-49	110047	93951	99300	115600	-11%	19%	5%
50-54	91125	77748	98700	87500	8%	11%	9%
55-59	63560	54303	73400	54900	13%	1%	8%
60-64	52202	44403	67400	36300	23%	-22%	7%
65-69	22920	19618	28000	18100	18%	-8%	8%
70-74	19101	16310	26100	16100	27%	-1%	16%
75-80	25356	20662	28700	20100	12%	-3%	6%
Total	2244340	1903963	2219500	2020400	-1%	6%	2%

The experiments design for population growth and standards' optimisation considered the two factors of fertility and migration with three levels of intensity as high, mid and low. This design resulted in nine alternatives to choose from. See Table 43 for projected total population counts for year 2070 under different scenarios of fertility and migration using original ADA population estimates.

Table 43 projected total population counts for year 2070 under different scenarios of fertility and migration using original ADA population estimates.

Factors		Migration		
		High	Mid	Low
Fertility	High	35,892,872	27,631,367	24,217,763
	Mid	24,465,437	18,864,269	16,538,915
	Low	19,690,025	15,259,532	13,408,184

Table 43 shows that for the period between 2015 and 2070, Riyadh city population is expected to grow from 4.15 million to 13.41 under the most limiting assumptions. The net growth for this scenario is equivalent to 223%. On the contrary, for the highest expected rates of fertility and migration, the city population is expected to have a net growth of 765% pushing the city population to around of 35.89 million. However, the author believes that mid and high fertility rates are unlikely to happen given the shift in population lifestyle towards a western social life style especially in metropolitan areas. Moreover, it is not expected for migration rates to drop in metropolitan areas with high levels of attraction (importance) such as Riyadh. Therefore, almost certainly lower fertility rates with mid or high migration rates are the more expected scenarios**.** The low level of fertility with mid or high migration rate to Riyadh city would set minimum expected population between 15.26 million and 19.70 million. Under the same circumstances of low fertility with mid and high migration rates the adjusted population is projected to be 16.00 million and 20.66 million respectively. The difference in projected population between original and adjusted base year data is around 5% in total. The effect of this increase on the targeted school age population will be explained in the coming subsection.

5.1.3 Schooling population

Regarding the targeted schools age population by educational stage and gender, first for the original population estimates, the total number of students for both genders in elementary, middle and secondary stages are estimated at 531858, 285490 and 250500 respectively. Moreover, the gender distribution is almost the same for all stages at 45% males and 55% females. When the original population estimates are projected, with low fertility and mid (current) migration rates, for the next 50 years, the number of students in elementary, middle and secondary stages are forecast to be at 1810032, 764587 and 554206 respectively. The growth in demand based on current capacity for each stage is expected to be around 340% for elementary, 268% for middle and 221% for secondary. However, if the migration rate is anticipated to double during this period then, the demand for each stage will increase roughly by 23%. Moreover, while, the gender gap is expected to get closer to zero over the first 15 years for elementary and middle stages, secondary education is expected to remain male biased by a marginal ratio of 1%. See Figure 38 for a projection of the original School age population by stage, gender and year for the next 50 years using low fertility and high migration rates.

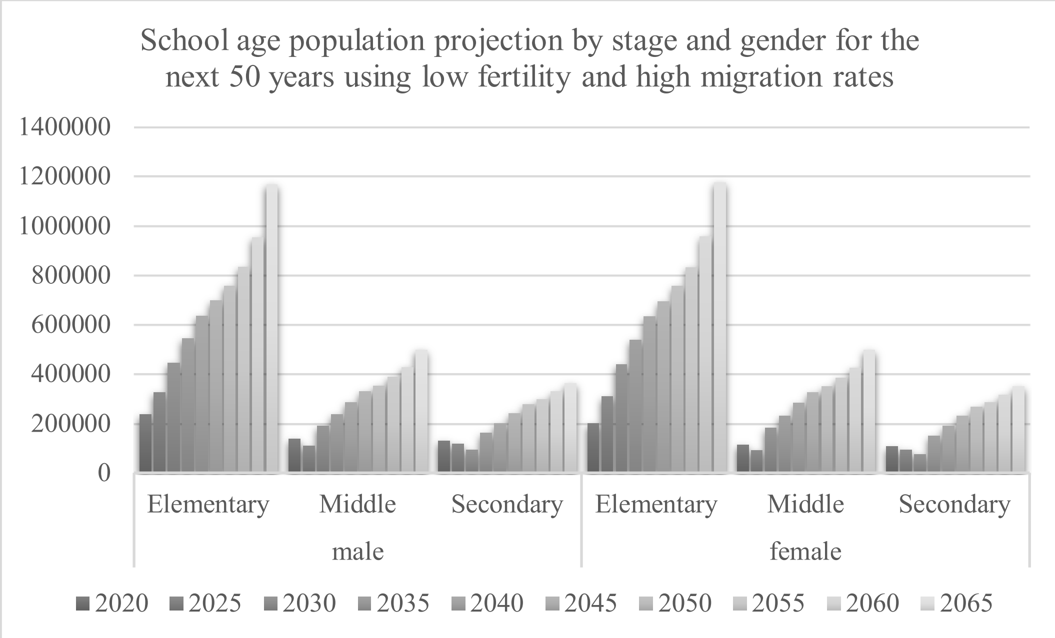


Figure 38 Projection of original School age population by stage, gender and year for the next 50 years using low fertility and high migration rates.

Figure 38 shows that the general trend for all education stages is expected to increase over the coming years. However there is a noticeable sudden drop in middle and secondary stages due to what is believed to be an underestimation of the number of births in the initial year. This drop is believed to have some effect on the number of future schools age population, when the females go through childbearing years. However the data available are lacking to sort out this issue similar to that done at the kingdom level in subsection C.2 Data curation methods.

When comparing the total school age population of adjusted base year population against the original population estimates, both the elementary and middle stages have negligible differences of (-1%) -4422 and (2%) 6050 students respectively. However, for secondary stage, there is around a 9% increase which accounts for 23946 students to be added to the demand side of the needed area for schools. Second, when the data is considered based on gender for elementary and middle stages, bigger differences can be seen with an average of (+-5%). Moreover, in elementary and middle stages the number of male students is underestimated and vice versa for females. See Table 44 for the difference in students' numbers by stage and gender using original and adjusted population data for further insights.

Table 44 the difference in students' numbers by stage and gender using original and adjusted population data..

Base year data	Stage	Male	Female	Total
Original estimates	Elementary	288839.7	243018.4	531858.2
	Middle	155735.4	129754.2	285489.7
	Secondary	137761.1	112738.4	250499.5
Adjusted estimates (Difference percentage)	Elementary	271185.6 (-7%)	256250.4 (5%)	527436 (-1%)
	Middle	151833.6 (-3%)	139707 (7%)	291540.6 (2%)
	Secondary	150494.8 (8%)	123951.4 (9%)	274446.1 (9%)

Despite that total difference in population among original and adjusted data is around 5%, the projection of adjusted data with low fertility and mid migration levels causes the school age population to additionally increase over the original estimated by 18% for elementary, 13% for middle and 11% for secondary. This increase is caused by the inflation of female population in the adjusted population mentioned earlier. The effect of higher level of migration results in approximately a 5% increase in school age population. See Table 45 for the projection of adjusted School age population by stage, gender and year for the next 50 years using low fertility and high migration rates.

Table 45 Projection of adjusted School age population by stage, gender and year for the next 50 years using low fertility and high migration rates.

Year	Male			Female		
	Elementary	**Middle**	**Secondary**	**Elementary**	**Middle**	**Secondary**
2020	213093	126824	127570	203875	119769	114242
2025	332684	97482	107414	324228	92483	98703
2030	467833	200478	83788	462492	194231	77376
2035	572061	249099	169267	570622	244784	159597
2040	672831	302224	209986	672628	299079	200807
2045	739187	350683	254484	739149	347164	245077
2050	803994	376384	295200	803716	372721	284426
2055	878196	412486	317294	879888	408290	305831
2060	1000000	450652	348186	1007000	447679	335508
2065	1230000	523158	381014	1240000	522616	368493

Based on the nine alternatives three scenarios were used for the standard optimisation as described in the coming section.

5.1.4 Optimised standards:

The optimisation at the city level will include the following subsections the considered scenarios, the optimisation objectives and variables, the assessment criteria, the results of the three most likely scenarios to occur and the city level results remarks.

5.1.4.1 Considered scenarios

Three of the nine feasible optimisation solutions were chosen for discussion. These three scenarios are as follows: low fertility with moderate migration, low fertility with substantial migration, and high fertility with high migration rates. The first two are seen to be more realistic, as previously stated. As a preventative measure, the very high end of the optimisation spectrum was included to understand what is required in the worst-case situation. Despite the fact that the upper end is unlikely to occur, the author believes it is the best choice to pursue. This is due to the fact that it not only serves the aim of optimising the dynamics of land supply and demand, but it also aids in planning for the worst-case situation in a foreseen glimpse of the future.

5.1.4.2 Optimisation objectives and variables

The major goal of the process at the city level is to keep the disparity between available and necessary land area as small as possible for each iteration over the next 50 years. The available land area is determined for each iteration by applying a restricted land consumption ratio (also known as the land preservation ratio) to the total available land area. For each educational level, the needed land area is determined by multiplying the number of pupils by the new school provision standards recommended. To achieve the ultimate goal, each iteration's objectives must be zero to achieve the ultimate goal. Furthermore, if zero difference is not achieved, the best viable solution is skewed toward having a surplus of land supply. As a result, 13 variables are included in the optimisation procedure. With no distinction between genders, three parameters cover the supply of each school stage. After that, for each time iteration, a land consumption parameter is added. The school provisioning requirements were controlled by global ratios, and the land consumption parameters were simply set to a maximum of 30% for each iteration.

5.1.4.3 Assessment criteria

The optimisation results consist of three additional elements to be highlighted beside the optimised standards of each education stage. These additional aspects include 1) the land preservation ratios; 2) the total amount of white land available for educational purposes;

and 3) the score of the optimisation objective which indicates any over or under supply of the total education area.

5.1.4.4 The first scenario (low fertility – current migration)

For the original population estimates with low fertility and mid (current) migration rates, the optimised provision standards (m^2/ student) for elementary, middle and secondary stages were at 3.90, 2.70 and 1.52 respectively. Moreover, the land preservation ratio averaged 92% which is equivalent to consuming (8% ± 5%) on average for every five years of the education land stock. Consequently, for the optimisation time horizon, it is estimated that the area of total consumed land would be 7,541,787 m^2, and the remaining area of land stock is at 6,116,000 m^2. Furthermore, the optimisation objective averaged (442 m^2 ± 4138 m^2). However, in most cases, it is expected to have a surplus amount of land associated with the use of the land preservation ratios. In years 2030, 2040 and 2050 and 2055 some insignificant shortage is expected in total area needed at -6501, -4770, -1279 and -3509 respectively. These shortages add up to -16059 m^2 in total which could be easily covered using the remaining land stock. Nevertheless, if the shortage has been accepted to as happening, it is not expected to cause any noticeable overcrowding in schools given the scale of current schools' capacity.

If the adjusted base year population estimates are used with the same parameters calibrated for the original base year population estimates, then the total amount of needed area will increase by 7.70% to 8,189,558 m^2. This implementation will cause an almost continuous shortage of supply that will accumulate to -661830 m^2 over the next 50 years. However, covering the land shortage from the available land stock will set the land stock to be at 5,468,170 m^2 which means consuming about 60% of the current (2015) land stock. Alternatively, adjusting the provision standards (i.e. calibrating for the adjusted population) to reduce the amount of needed land is a feasible solution. The recalibration of the standards resulted in the provision standards (m^2/ student) for elementary, middle and secondary stages to be 3.214, 2.592 and 1.591 respectively. This means that changing the standards by -17.55 %, -4.0% and +4.53% respectively, would bring the total needed area to 7,128,787 m^2. Thus, the total amount of land stock will be at higher levels without major influence on the standards at least for the middle and secondary stages. See Table 46 for a summary of standards optimisation outcomes using low fertility with current migration ratios.

Table 46 Summary of standards optimisation outcomes using low fertility with current migration ratios.

Year	Parameter	Original Opti. parameters	Adjusted Opti. parameters	Land stock*	Needed Land area*	Original opti. objectives	Adjusted opti. objectives based on original	Adjusted opti. objectives
All	Elementary	3.898	3.214					
	Middle	2.7	2.592					
	Secondary	1.522	1.591					
2020	Stock land preservation ratio	99.80%	100%	1364	1.78	5399	10299	9528
2025		94.70%	95%	1292	72.29	839	-153772	31
2030		90.00%	90%	1163	129.2	-6501	-187875	87
2035		91.00%	91%	1061	101.4	3135	-82146	32554
2040		92.00%	92%	979.3	81.84	-4770	-81431	30537
2045		94.50%	95%	925.5	53.86	4171	-58425	25
2050		95.90%	96%	889.5	35.94	-1279	-25659	20057
2055		95.10%	95%	853.3	36.2	-3509	17503	73839
2060		91.10%	91%	790.5	62.86	908	14396	130864
2065		79.90%	80%	652.9	137.6	6029	-72523	213109
*	10,000s							

5.1.4.5 The second scenario (low fertility – high migration)

Moving into population projections with low fertility and high migration rates, for the original population estimates with low fertility and high (double) migration rates, the optimised provision standards (m2/ student) for elementary, middle and secondary stages were at 3.82, 2.067 and 1.635 respectively. Moreover, the land preservation ratio averaged 89% which is equivalent to consuming (11% ± 9%) on average for every five years of the education land stock. Consequently, for the optimisation time horizon, it is estimated that the area of total consumed land would be 9,826,787 m^2, and the remaining area of land stock at 3,831,000 m^2. Furthermore, the optimisation objective averaged (589m2 ± 3307 m^2). The amount of shortage in land associated with the use of preservation ratios is insignificant. The accumulated total of shortage is -11220 m^2 which could be simply covered using the available land stock. If a slight shortage is accepted to happen, then noticeable overcrowding in schools will not be an issue given the scale of current schools' capacity.

If the adjusted base year population estimates are used with the same parameters calibrated for the original base year population estimates, then the total amount of needed area will increase by 7.14% to 10,593,967 m^2. This implementation will cause an almost continuous shortage of supply that will accumulate to -758176 m^2 over the next 50 years. However, covering the land shortage from the available land stock will set the land stock at 3,064,824 m^2 which means consuming about 78% of the current (2015) land stock.

Alternatively, adjusting the provision standards (i.e. calibrating for the adjusted population) to reduce the amount of needed land is a feasible solution, the recalibration of the standards resulted in provision standards (m^2/ student) for elementary, middle and secondary stages to be 3.64, 2.503 and 1.688 respectively. This means changing the standards by -4.95 %, +17.42 % and +3.14% respectively, which would bring the total needed area to 10,628,107 m^2. Thus, the total amount of land stock will be less than that produced by the original data optimisation. See Table 47 for a summary of standards optimisation outcomes using low fertility with high migration ratios.

Table 47 Summary of standards optimisation outcomes using low fertility with high migration ratios.

Year	Parameter	Original Opti. parameters	Adjusted Opti. parameters	Land stock*	Needed Land area*	Original opti. objectives	Adjusted opti. objectives based on original	Adjusted opti. objectives
All	Elementary	3.82	3.64					
	Middle	2.067	2.503					
	Secondary	1.635	1.688					
2020	Stock land preservation ratio	99.80%	99.80%	1364	2.10	1723	6996	6337.562
2025		94.40%	93.60%	1276	87.34	162	-152887	-639.062
2030		89.90%	88.20%	1126	150.80	-2047	-167402	-1548.31
2035		89.70%	88.40%	995.4	130.40	-1219	-98319	1620.104
2040		89.90%	88.50%	880.9	114.50	3828	-82079	-619.818
2045		91.70%	90.10%	793.7	87.52	-2606	-77659	-3086.534
2050		92.00%	90.70%	720.2	73.44	-3667	-46404	3755.031
2055		89.90%	88.90%	640.3	80.02	4817	1531	-716.434
2060		83.20%	81.20%	519.9	120.40	-1681	-22236	-47.455
2065		64.90%	58.30%	303.6	216.30	6575	-111191	4787.15
*	10,000s							

5.1.4.6 The third scenario (high fertility – high migration)

For the highest anticipated levels of fertility and migration with original population estimates, the optimised provision standards (m^2/ student) for elementary, middle and secondary stages were at 1.256, 1.255 and 1.28 respectively. Moreover, land preservation ratio averaged 78% which is equivalent to consuming (13%, ± 14%) on average for every five years of the education land stock. Consequently, for the optimisation time horizon, it is estimated that the area of total consumed land would be 10,624,787 m^2, and the remaining area of land stock at 3,033,000 m^2. Furthermore, the optimisation objective averaged (327m2 ± 5713 m^2). The amount of shortage in land associated with the use of preservation ratios is also inconsequential. The accumulated total of shortage is -21811 m^2 which could be simply covered using the available land stock. If a slight shortage is accepted to happen then, noticeable overcrowding in schools will not be an issue given the scale of current schools' capacity.

If the adjusted base year population estimates are used with the same parameters calibrated for the original base year population estimates, then, the total amount of needed area will increase by 6.30% to 11,316,618 m^2. This implementation will cause a continuous shortage of supply that will accumulate to -676,212 m^2 over the next 50 years. However, covering the land shortage from the available land stock will set the land stock to be at 2,340,788 m^2 which means consuming about 83% of the current (2015) land stock. Alternatively, adjusting the provision standards (i.e. calibrating for the adjusted population) to reduce the amount of needed land is a feasible solution. The recalibration of the standards resulted in the provision standards (m^2/ student) for elementary, middle and secondary stages to be 1.242, 1.204 and 1.097 respectively. This means changing the standards by -1.13 %, -4.24 % and -16.68% respectively, would bringing the total needed area to 10,889,878m^2. Opposite to all previous optimisations for adjusted population estimates, the elementary stage had the lowest percentage of standards decrease. Moreover, no gains in any standards were observed. See Table 48 for a summary of standards optimisation outcomes using high fertility with high migration ratios.

Table 48 Summary of standards optimisation outcomes using high fertility with high migration ratios.

Year	Parameter	Original Opti. parameters	Adjusted Opti. parameters	Land stock*	Needed Land area*	Original opti. objectives	Adjusted opti. objectives based on original	Adjusted opti. objectives
All	Elementary	1.256	1.242					
	Middle	1.255	1.204					
	Secondary	1.28	1.097					
2020	Stock land preservation ratio	99.90%	99.90%	1,364.0	1.38	-6,535	-2,415	-117
2025		96.50%	96.10%	1,311.0	53.87	-5,101	-67,212	-6,563
2030		92.60%	92.10%	1,208.0	103.50	6,955	-85,103	1,044
2035		91.30%	91.10%	1,100.0	107.50	4,800	-81,552	326
2040		91.90%	91.50%	1,007.0	93.59	-3,695	-80,142	-652
2045		92.70%	92.30%	929.8	76.95	-1,632	-69,179	5,684
2050		93.20%	93.10%	865.8	64.00	-3,046	-33,183	1,524
2055		88.60%	88.40%	765.7	100.10	-1,802	-26,963	3,134
2060		76.60%	75.80%	580.8	184.90	768	-71,951	3,836
2065		99.90%	99.90%	1,364.0	1.38	-6,535	-2,415	-117
*	10,000s							

5.1.5Remarks on the city level optimisation

To conclude the standards of the three prospective trajectories, the differences between the adjusted values will be discussed. This is because it maximises the number of included students with a minor effect on the provision standards as shown earlier in the three optimisation scenarios, especially for the larger elementary population. Also, the adjusted population should be directly comparable to the outcomes of the district model in the next section. See Table 49 for the summary of optimised standards by adopted rates and education stages along with the total needed area by each scenario.

Table 49 The summary of optimised standards by adopted rates and education stages along with the total needed area by each scenario.

Aspect		Low fertility current migration	Low fertility high migration	High fertility high migration
Education stage	Elementary	3.214	3.64	1.242
	Middle	2.592	2.503	1.204
	Secondary	1.591	1.688	1.097
Total needed land area		7,129,709	10,628,107	10,889,878

Table 49 shows that both standards with low fertility rate seem to be identical. Although, opposite to the expectation of decreasing standards with a higher number of students, the scenario with higher levels of migration had slightly better standards. The difference in the standards between current and high migration rates for elementary, middle and secondary stages were +13.25%, -3.43% and +6.10% respectively. This highlights the ability to maintain a high level of provision standards with changing migration conditions. However, the amount of consumed land will be subject to a significant change. The differences between current and high migration on the amount of consumed education land is 52% and 78% respectively. Despite the significant increase in consumed land, the current land area in total is proven to be more than sufficient for population estimates with low fertility and high migration rates. If the decision maker is keen to preserve more land by adopting the lowest standards from both optimisations, then the amount of consumed land could be decreased to 71% with the high level of migration level. By taking the worst-case scenario of high fertility with high migration rates, the provision standards for elementary, middle and secondary stages will shrink by 65.88%, 51.90% and 35.01% respectively. However, this adverse change is within the acceptable levels of global supply standards. Moreover, these standards are higher than the current standards suggested by ADA. Furthermore, these rates will consume about 80% of the education land which is comparable to the amount of land consumed by the standards optimised for low fertility with high migration rates. Therefore, it could be argued that the current available land for education is more than adequate for the next 50 years. Also, the supply of schools at the highest standards eventually could absorb the increase of students resulting from booming population growth and the crowding level would probably be acceptable. Another consideration that could improve both the amount of preserved land and the provision standards would be possible by changing the building regulation FAR to increase the number of allowed school building stories.

Given these outcomes it becomes clear that there would be another problem associated with the distribution of this land within the city districts.

5.2 Districts level

This section will be split into the following subsections. First, the population growth of the ABM-based district level model is compared to that of the CCM-based city level model. The number of school-age children will then be summarised for the projected time horizon. Following that, the optimisation findings for three selected districts from Riyadh's 150 districts will be displayed. These three districts should cover a wide range of common optimising scenarios. Furthermore, in the next subsections, the optimisation outcomes of both parcel-specific and non-parcel-specific functions will be shown and described. The final remarks for the optimisation process will then be given.

5.2.1 Population growth

Given the modifications made to the traditional CCM model, it can now be used as an agent-based model. By default, the results of the two models are not expected to be similar. This is due to the fact that the idea of households is applied to the population, which introduces extra aspects into the modelling process, such as marriage and divorce. All the other variables that affect the model's outputs are discussed in depth earlier in chapter 4 Methodology. Despite the fact that these parameters may help us better understand the drivers of population change, the present goal of this ABM model is to reproduce the CCM results as being a standard benchmark model. As a result, with the inclusion of a weighting parameter, the ABM population estimates were calibrated to replicate the CCM outcomes. The extra parameter was introduced to the fertility rates largely to manage the new-born cohort's growth rates. The rate has only a multiplicative influence on fertility, and it brought the model's outputs closer together in a way that should be acceptable. Other key social factors may be used to improve the calibration procedure even further. However, due to a lack of data and competence in other areas, the risk of inaccurate calibration is increased and therefore not tested. The next paragraph will go through both the original and calibrated outcomes.

Another aspect influencing population predictions is the distribution of demographic characteristics among newly migrated people. The migrating population may be simply incorporated into the CCM technique by raising the population numbers based on the likelihood linked with the demographic characteristics. However, in order for the ABM model to work, all extra people must be added to a household's units. Ultimately, this issue necessitates the employment of a population synthesiser to ensure the conformity of additional household characteristics to actual migrating households. Given the absence of this function at this point, the adult population (20 and above) of both genders is added, followed by marriage and births to compensate for the missing 0–4-year-old migrating population. This method has the disadvantage of excluding the population aged 5 to 19, but it gives adequate closeness to the CCM results, as demonstrated in the next subsections. This could be justified by the literature on migrated populations, which supports accepting this approach, as the age group 5–19 is often linked with low migration likelihood.

5.2.1.1 Comparing CCM and ABM Population projections

When comparing ABM and CCM results, various fertility and migration levels must be taken into account. According to the data, overall population counts estimated using the CCM are 35% higher on average than those predicted using the ABM, with a standard deviation of (+/-8%). See Table 50 for the ABM population estimates and the percentage difference between them and the CCM results for various fertility and migration rates without fertility recalibration.

Table 50 the ABM population estimates and the percentage difference between them and the CCM results for various fertility and migration rates without fertility recalibration.

Factors		Migration		
		High	Mid	Low
		(% percentage difference to CCM)		
fertility	High	19,316,797 (-49%)	18,017,180 (-38%)	17,492,432 (-31%)
	Mid	14,621,971 (-43%)	13,477,046 (-32%)	12,929,610 (-26%)
	Low	11,861,423 (-43%)	10,871,694 (-32%)	10,465,024 (-26%)

This is understandable, as CCM population reproduction employs an age-specific fertility ratio for all women of reproductive age. However, in the ABM, population reproduction is influenced by marriage, divorce, and parent loss. To address this issue, the ABM approach reweights fertility based on a single variable in order to match the CCM population estimates. The ABM calibration revealed a nonlinear link between fertility reweighting and model outputs. In other words, increasing the ratio of various fertility rates does not have an equivalent effect on reweighted fertility outcomes. Calibration values were 1.45, 1.62, and 4.8 for the most likely scenarios of low fertility with low migration, low fertility with high migration, and high fertility with high migration, respectively. While the overall population counts are taken into consideration when assessing the calibration rates discussed before, it is discovered that the disparities are within a range of -5.65 % to +0.18 %, which is considered a permissible error. The difference in total population outcomes between calibrated ABM and adjusted CCM for the year 2070 is shown in Table 51.

Table 51 change difference between calibrated ABM and Adjusted CCM total population outcomes for year 2070.

Factors		Migration		
		High	Mid	Low
		(% percentage difference to CCM)		
fertility	High	37853210 (+0.18%)	-	-
	Mid	-	-	-
	Low	16149912 (+0.90%)	19497104 (-5.65%)	-

The expected error level for the calibration outcomes of all three scenarios is 14 % MAPE for all age groups. When we consider the population that is reproduced in the model via the birth function, the error level drops to an average of 7% MAPE. This case demonstrates the significant influence that tweaking the fertility parameter alone may have in modifying model outputs. Another intriguing feature was the underestimation of elderly men. Males aged 50 to 69 saw an average population reduction of 27% across all scenarios, compared to 10% for elderly females. Surprisingly, males over the age of 70 experienced a 49 % decline, while females in the same age range experienced a 21 % decline.

5.2.1.2 Schooling population

For the school-age demographic that is being targeted, the average error for the three cases is 11% MAPE with an 8 % standard deviation. As a result, the highest overestimation of students over the projection time horizon is slightly more than 170,000, while the maximum underestimate is 223,000 students. Although these figures appear to be significant, the demand they generate is expected to be divided across the 150 districts in Riyadh and should be manageable given the process of standardisation taken into account. Figure 39 depicts school-age population forecasts for the next 50 years based on CCM and ABM, assuming low fertility and high migration.

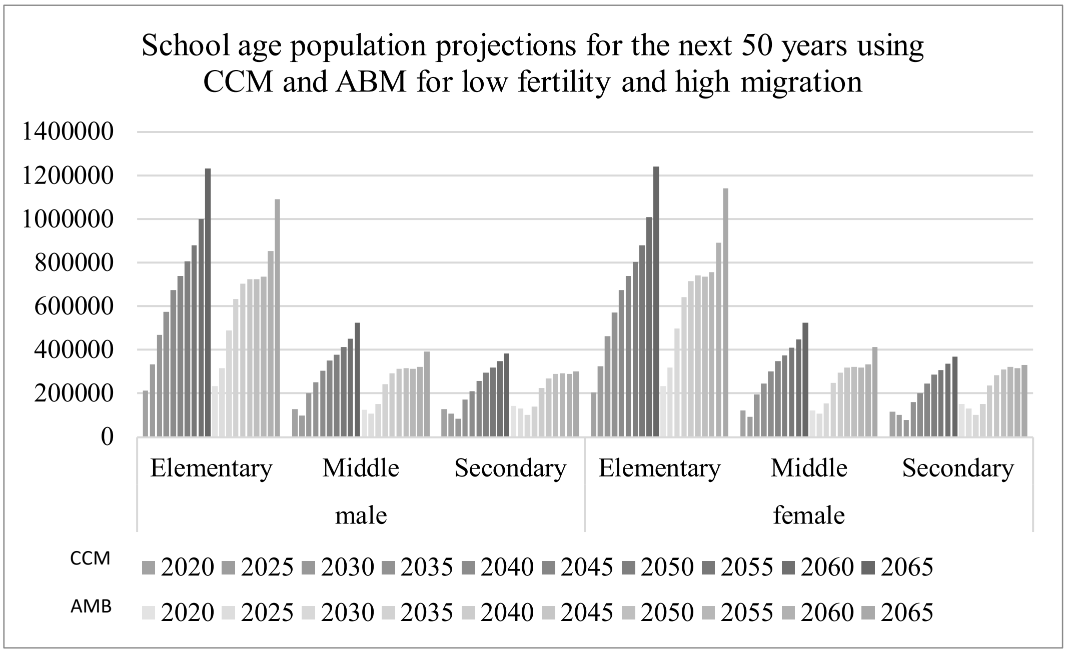


Figure 39 School age population projections for the next 50 years using CCM and ABM for low fertility and high migration.

Figure 39 demonstrates the strong similarities between the CCM and ABM outputs. It is also obvious that the CCM exhibits more stable growth patterns over time than the ABM random fluctuations. Furthermore, CCM has a higher expectation for population size, although not in a detrimental way. Despite considering any aspect such as education stage, gender and the level of fertility or migration the MAPE level remains very close to 11 % with standard deviation of 8%.

5.2.2 Optimised standard

5.2.2.1 Introduction

The discussion in this part will centre on the optimisation functions' ability to reach targeted supply levels given that the inputs will vary depending on the final user's assumptions for population change. Additionally, the results will be sorted by the degree of variance and specifics in each district, depending on the various situations ranging from low to high variance. This is because a sample of three districts was chosen to reflect the most likely scenarios among Riyadh's 150 districts. These districts were chosen based on their area size and population growth pattern, which will be discussed in further depth in the next part.

The district-level optimisation process differs from the city-level optimisation method. This is founded on three factors. First, criteria such as square metres per student and land consumption ratios are optimised at the city level. At the district level, an effort is made to become more practical by placing MOE predesigned school building prototypes on lands reserved for educational reasons. As a result, the optimisation solution is created by selecting predesigned school combinations that will be converted into counts of additional education capabilities by stage and gender.

Second, in contrast to city-level optimisation, which aims to reduce each time-iteration (5 years) separately yet concurrently. The optimisation at the district level considers the entire time span (50 years) as a single task. This is due to the substantial increase in the number of factors required to optimise the district level using 10 separate time iterations. As a result, determining an average demand level for the next 50 years was a critical component of this strategy.

Third, based on the size of the district's demand, the district level optimisation has two separate methods to choose from. The parcel-specific function should be utilised in circumstances of normal or low demand, while the non-parcel specific model should be used in cases of high demand or unknown land constraints. Regardless of the quantity of lands available, both parcel specific and non-parcel specific optimisation methods were evaluated on all three districts for demonstration purposes. The restriction on land size has been eliminated for districts with land parcels greater than ten. Also, to have a better knowledge of the behaviour of the model, for each case, the number of optimisation iterations was set at 200,000. This is because, after around 200.000 iterations, the optimiser tends to become extremely slow in finding a better solution, which is a time-consuming operation. Also, it is worth mentioning that for the non-parcel specific optimisation, each school type was limited to 50 units which is equivalent to 2200 schools in total. This is mainly because leaving the number of schools unlimited will lead to selecting the schools with the highest capacities in high numbers and ignoring the schools with smaller capacities. Also, the upper limits of 50 schools for each type or a total of 2200 schools is significant and unreasonable for typical size district.

However, all functions have the same limitation of not allowing current supply reduction. Regardless of the differences in the parameters and objectives evaluated by optimisation functions at the city and district levels, the ultimate goal is the same: to optimise the location and distribution of educational land.

It is critical to define the terminologies used in this section at this point. To begin, the actual number of students refers to the number of students estimated by the simulation model of population change. The actual cumulative moving average (actual CMA) refers to averaging the actual numbers of students over 10 iterations in a 50-year time horizon using a cumulative moving average approach. Third, the targeted average for the simulation process, also called as the targeted cumulative moving average (targeted CMA), is the result of averaging all the actual CMA data into a single value. Thus, the actual student numbers have been averaged twice: first to smooth the actual data, and then to calculate the average of the smoothed data, which has been considered as a target for optimising supply and demand.

The method of selecting the three sample districts to combine with the previously established most likely population growth scenarios is described in the following section.

5.2.2.2 Considered scenarios

Along with the preselected most likely population growth scenarios of low-fertility with mid- migration, low-fertility with high-migration and high levels of fertility and migration, additional district-related aspects must be included. The factors included in the analysis are the trend of change in students counts over time at the district level, the count of available educational lands and the size of the district and how typical it is among Riyadh city districts.

To define the scenarios that can be considered as a result of each of these components, first, the number of students in each district can be summarised as 1) constantly increasing; 2) constantly dropping; 3) fluctuationally increasing; 4) fluctuationally decreasing; and 5) relatively steady, depending on the residential mobility model used.

Second, the average number of land parcels dedicated for education purposes among Riyadh city district is 6.7 land parcels with standard deviation of 7.8 land parcels. However, on the one hand, some districts may be short on educational land, prohibiting the optimisation from accomplishing its goal. On the other hand, some districts may have a large number of available education land parcels, which substantially broadening the space of optimisation alternatives

Third, due to the disparities in district sizes, a direct comparison of the districts' demands would not be suitable, as would the use of direct comparison of population density. Therefore, an additional layer of classification based on how similar the districts' characteristics would be useful practice. The author recognises two main patterns based on the districts' sizes in Riyadh districts which are the typical and non-typical districts. In terms of the criterion that determines whether a district is typical or unique, districts with an area of less than 8 km^2 are considered typical. This is due to the fact that in Riyadh city, standard planning units for districts are 4 km^2 in size, such as district 124, and it is common to have two adjoining planning units considered one district, such as districts 63 and 111. Additionally, the organically developed districts in the old city centre are considered typical and normally have a small area size, as in district 140. The unique districts, on the other hand, are mostly those areas that are located on the city's present outskirts such as districts 139, 19, 15 and 53. These districts are either divided into land subdivisions with soft boundaries known to locals and the land market or they are still undivided. See Figure 40 for the location of the initial sample districts in Riyadh city labelled by their code. Also, see Table 52 for the characteristics of the initial sample of considered districts for evaluating the district level optimisation performance.

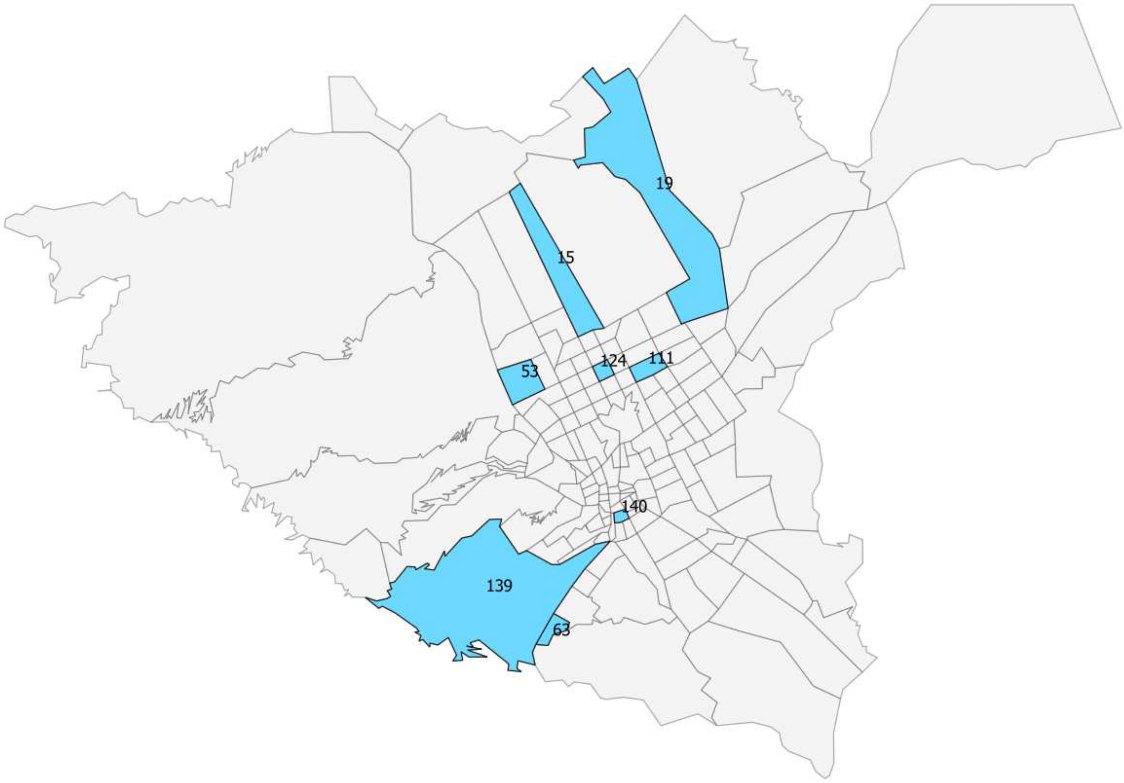


Figure 40 the location of the initial sample districts in Riyadh city labelled by their code.

Figure 40 shows the location of the initially sampled districts in Riyadh city. Also, the difference in area size between the standard area of district 124 and the non-standard districts, which cover enormous areas in most cases.

Table 52 the characteristics of the initial sample of considered districts for evaluating the district level optimisation performance.

Distinctive type	District ID	District area (Km^2)	Count of total education white lands	Area of total education white land (m^2)	projection trend of population*		
					(SI)	(NN)	(Hy)
Unique	139	244	59	682826	CIT	FDT	CIT
	19	173.9	0	0	FIT	FIT	CIT
	15	49	8	115300	CIT	FDT	CIT
	53	18.4	3	32584	FDT	FIT	FIT
Typical	111	8.3	21	144654	FDT	FIT	FIT
	63	6.9	1	7894	FST	CDT	FIT
	124	4	7	109071	FDT	FIT	FIT
	140	2.1	3	17033	FDT	FIT	CIT

CIT **Continuous increasing trend**
CDT **Continuous decreasing trend**
FIT **Fluctuant increasing trend**
FDT **Fluctuant decreasing trend**
FST **Fluctuant but stable trend**
*** Population trends reflects on both genders in the secondary education stage**

Table 52 summarises eight districts by their distinct type, code ID, area (Km^2), number of available education white lands, total area of available education white lands, and a comprehensive description of their school age population forecasted by the various mobility models. The districts sample demonstrates a range of possible outcomes in Riyadh's districts, including large districts with diverse amounts of land dedicated to educational purposes and additionally, how the trend in student counts varies according to the different residential mobility options. For instance, district 140 has a distinct trend for each residential mobility method.

The three districts chosen were 124, 139 and 140. This is because district 124 represents the most typical planning unit in terms of its size; its count of available land for educational purposes equivalent to the overall average, and it has both increasing and decreasing trends in population change. Also, district 124 represents many of the almost completely occupied districts in Riyadh city. Then, district 139 was chosen because it has the largest district area, has the most available land for educational purposes, and has both main trends of population change. In contrast to district 124, district 139 has a vast amount of residential white land to be developed and a huge potential for population growth given the current trend of development. District 140 was chosen to represent the districts of the old city which are normally considered for gentrification due to the outflow of residents, small in area size and has reasonable amount of land for education purposes relative to its area.

5.2.2.3 Optimisation objectives and variables

The main goal for the district level optimisation functions is to achieve an average level of supply for the anticipated coming 50 years. The targeted supply level is defined as the overall average of the smoothed actual students number projected by the population change model. The targeted supply is achieved by varying combinations of predesigned school buildings (i.e. school prototypes) for all schools by stage and gender simultaneously. The original number of building prototypes provided by the MOE is 22. However, due to the potential of being able to construct more than one building in large land parcels, the 22 prototypes were combined to create 184 different combinations of non-duplicated school stages. This step should maximise the use of land parcels beyond considering one education stage per parcel. Nevertheless, the requirement of 3 parameters per land parcel defines the utilised prototype ID, the gender of the school and the size of the land parcel. The number of parameters becomes an obstacle for districts with a large amount of land dedicated to education purposes. Therefore, two separate but similar functions were developed based on the number of land parcels being optimised. The parcel specific optimisation function is able to optimise for up to 10 land parcels. However, due to the possibility of suggesting multiple education stages, each land parcel can be divided theoretically in up to 3 parcels. On the other hand, the non-parcel specific model is designed to suggest the required number of any of the 22 building prototypes by gender. Consequently, the land related limitation is removed and its dependent on the number of buildings required results in a total of 44 parameters. This is because both genders were considered separately. Finally, both functions could be further constrained by total area used and the total counts of buildings suggested.

5.2.2.4 Assessment criteria

The optimisation results can be assessed based on the MAPE between the targeted average of students and supplied level to additional schooling capacities. The results also define how does the available land counts and areas compare to the optimisation suggestions.

5.2.3 Scenarios

The results shown in this section were organised according to the amount of variance exhibited in the results for each district. Therefore, the results of district 140 were first demonstrated, given the anticipated continuous oversupply of educational facilities that does not trigger the optimisation to alter the current supply levels. Then, this is followed by district 124, which required optimising the supply of male students in the secondary stage. Last, the optimisation outcomes for district 139 were explained due to the huge requirement for supply for all educational stages. Moreover, given the outstanding need for supply from the students' estimates generated by both SI and hybrid residential mobility models, these ranged from 527 to 1604 schools in total for all stages. Only the secondary stage students' estimated with the NN residential mobility were extensively elaborated upon. This is because of the manageable size of the needed schools, which reach up to 14 schools in total for all stages. The results were also compared to the ADA action plan for providing public facilities. Nevertheless, the time horizon of the ADA was limited to 2030. Also, the population estimates accounts for the non-Saudi citizens. As a consequence, it is impractical to compare the findings directly, yet it is being done so for reference purposes.

5.2.3.1 District 140

District 140 is one of the old districts in Riyadh city which has a total area of $2.1km^2$. Moreover, three land parcels are available for further supply of schools with a total area of 10220 m^2 with an average of $5677m^2$ and a standard deviation of $468m^2$. Regarding the population change, it is expected to increase under the residential mobility models of NN and hybrid. However, a light population decrease was expected under the SI mobility model. However, this is despite the mobility model being used for the current supply levels for all education stages which are beyond the anticipated demand. Therefore, both optimisation functions did not alter the supply level.

A. Elementary stage supply optimisation for district 140

The current elementary school supply level for each gender is roughly 16600 seats. Given that the average number of students expected in all scenarios is much lower than the current supply, which stands at 1102 students at most, both optimisation functions did not increase supply above the current levels. See Figure 41 for results of District 140 elementary stage supply optimisation. The ADA action plan, on the other hand, proposes expanding the number of primary schools for male students by 10, 1, and 1 school accordingly in the years 2016, 2020, and 2030. For female students, it is proposed that 10 schools be added in 2016 and one school be built in 2020. As a result, a total area of 184,825m2 is required to meet the ADA's expected need through 2030. The author believes that the increased demand predicted in the ADA work is being driven by the non-citizen population, which primarily lives in the city's older districts.

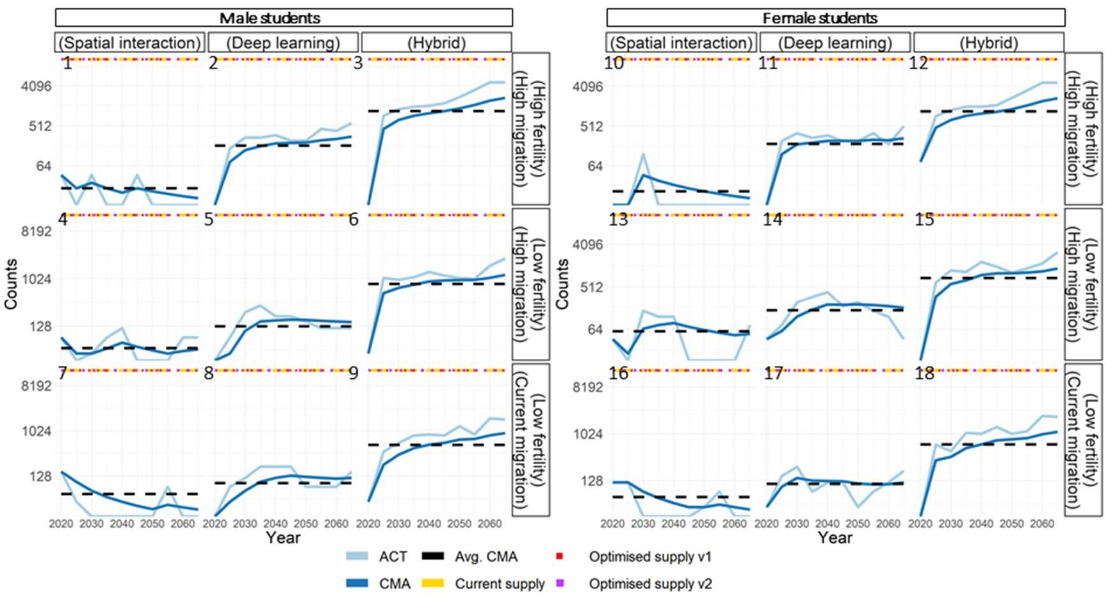


Figure 41 Results of District 140 elementary stage supply optimisation.

B. Middle stage supply optimisation for district 140

The current supply of middle school seats for each gender is approximately 6000. Given that the targeted average number of students in all situations is less than 360, neither of the optimisation functions increased supply above existing levels. See Figure 42 for results of District 140 middle stage supply optimisation. Similar to the elementary stage, the ADA action plan proposes the addition of five schools, one for each gender, totalling 106,000 m2, which would primarily serve non-Saudi students.

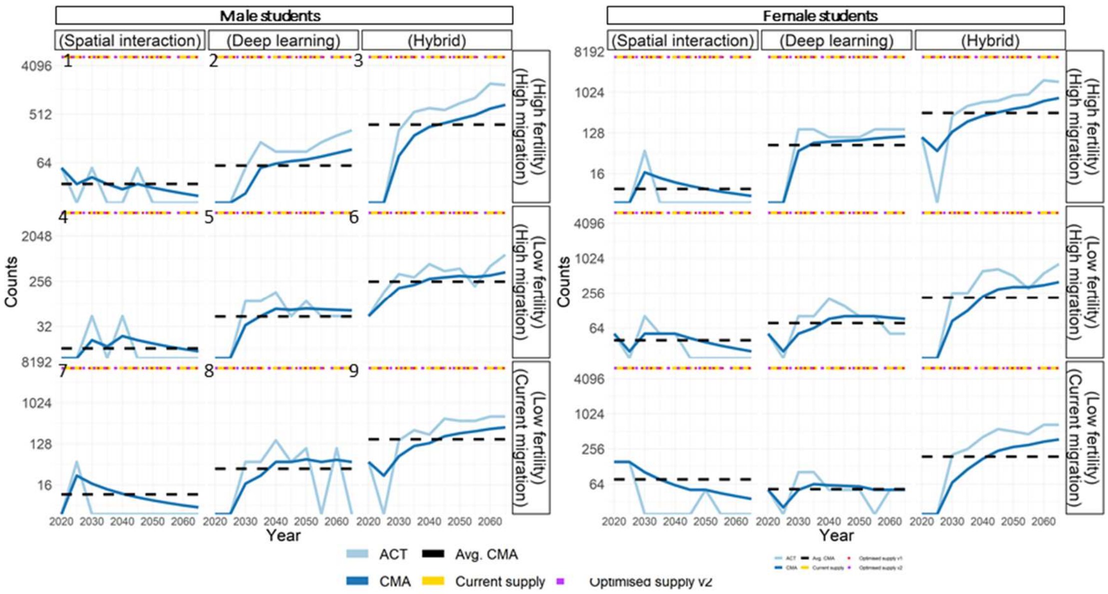


Figure 42 Results of District 140 middle stage supply optimisation.

C. Secondary stage supply optimisation for district 140

The current supply of secondary school seats for male and female students is 5406 and 3224, respectively. Given that the targeted average number of students in all situations is less than 415 for male students and 240 for female students. Neither optimisation function raised supply above current levels. See Figure 43 for results of District 140 secondary stage supply optimisation. The ADA action plan, like the elementary and middle phases, proposes the addition of 5 male and 4 female schools with a total size of 110000m^2 that would largely serve non-Saudi students.

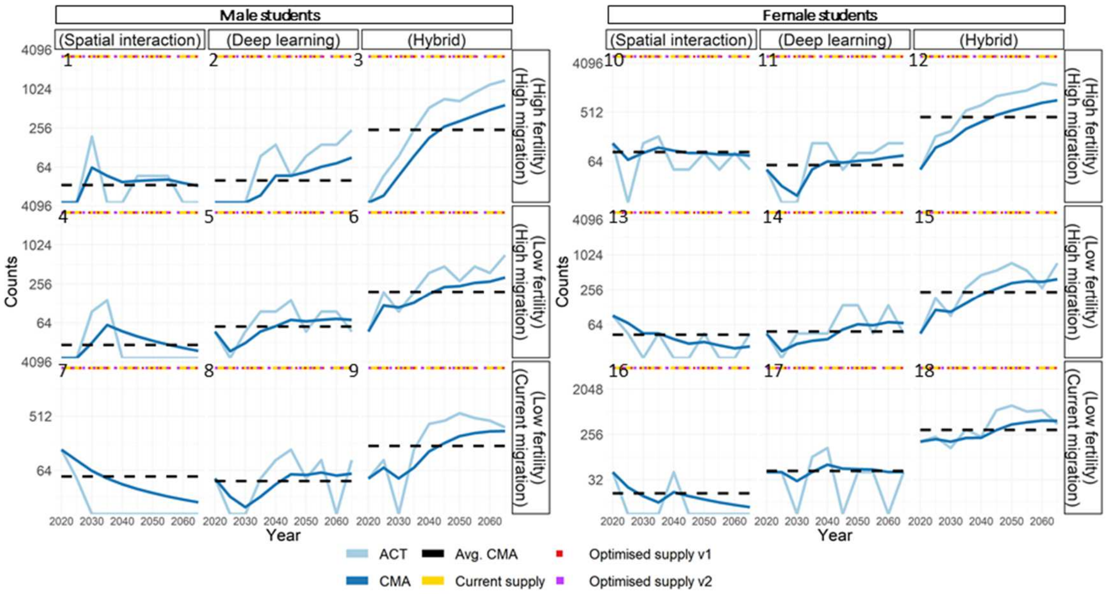


Figure 43 Results of District 140 secondary stage supply optimisation.

District 140 is a case of educational facilities being oversupplied for the targeted Saudi population. Additionally, it demonstrates how the optimisation function was sound in that it did not alter the supply in all circumstances. Additionally, the population change model provided a range and target average for adjusting current supply to the Saudi population. To accommodate non-citizen students' supply and demand, the entire process must be calibrated and run using their relevant inputs.

5.2.3.2 District 124

District 124 has a total size of 4 km^2, which is similar to a standard planning unit for Riyadh districts. In addition, seven pieces of land are still available for the construction of new educational buildings. The total area of these land parcels is 109070 m^2, with an average of 15500 m^2 and a standard deviation of 5170 m^2. Furthermore, based on the various demographic estimates for District 124, one may claim that the student population is projected to remain relatively steady in the foreseeable future. However, changes in fertility and migration levels could translate into increased demand in the distant future, although this can only be confirmed as time progresses until the early 2050s. Except for female secondary stage education, the district's present supply is often more than the cumulative demand averages. The optimisation findings for each stage are as follows:

A. Elementary stage supply optimisation for district 124

The present supply level for primary schooling for both genders is a little below 2000 seats. Given that the average number of students projected in all scenarios is less than the present supply, both optimisation functions did not raise supply over the existing levels. See Figure 44 for results of District 124 elementary stage supply optimisation. The ADA 2016 action plan, on the other hand, proposes expanding the number of primary schools for male and female students by 5 and 4 schools, respectively, with a total area of 72300m^2 to cover the need up to year 2030. This proposed increase goes against our estimation of sufficient supply.

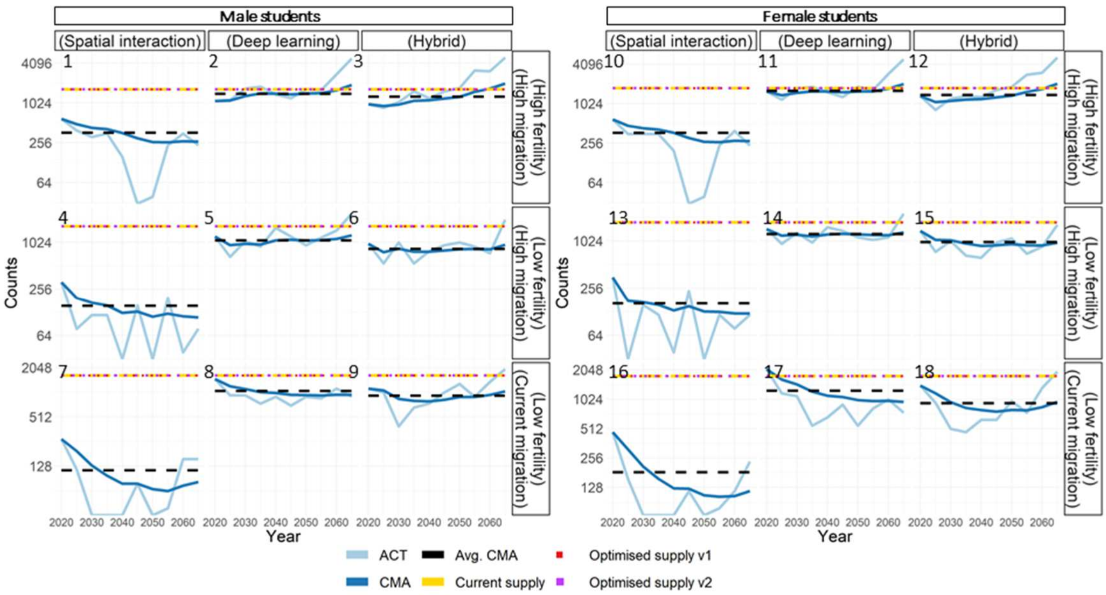


Figure 44 Results of District 124 elementary stage supply optimisation.

B. Middle stage supply optimisation for district 124

The current supply level for the middle stage is about 500 seats for male students and 1,000 seats for female students. In most scenarios, both supply levels outnumber the targeted number of students for an extended period of time. However, given the lower level of supply in male seats, the difference between the desired average number of students and the supply level is small, and in three situations (5, 8 and 9), current supply levels are lower than the initial number of students. Fortunately, cases of greater demand in the first 5 years are expected to fall significantly and are not imposing a worrying demand that would necessitate more supply to meet the expected number of average students. For example, in Case 8, the average demand is just slightly higher than the current supply mark, which is insignificant for the optimiser to alter the current supply.

Only in the late 2040s might it be expected that the average number of students would necessitate some supply adjustment, as in cases (2, 3, 11 and 12). However, this issue should be evaluated in the future. Until now, the optimisation function has not acted in a way that is intended to boost demand. Figure 45 show the results of District 124 middle stage supply optimisation.

According to the ADA, three male and two female schools with a total space of 62800m2 are required to meet demand through 2030. Despite the fact that the ADA results recommend more school supplies, our findings advise against it. On the other hand, the greater supply for boys' education can be justified by its current lower of supply.

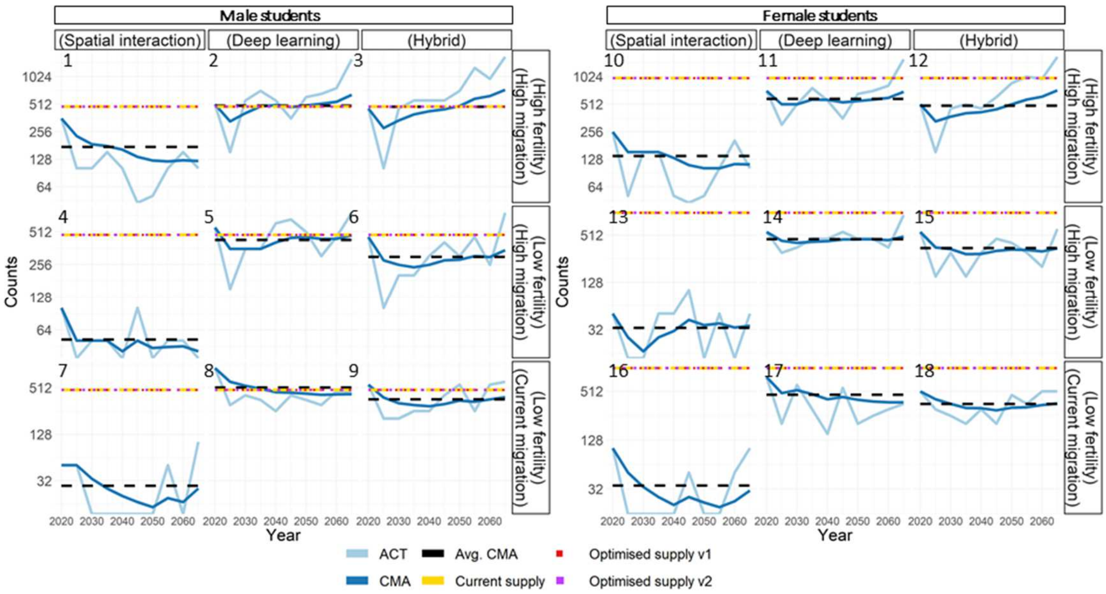


Figure 45 results of District 124 middle stage supply optimisation

C. Secondary stage supply optimisation for district 124

Regarding the current levels of secondary education provision, there is a significant discrepancy in supply, with males having zero seats and females having 958 seats. Furthermore, it is obvious that both optimisation functions behaved identically to the shortfall. For male students, the projections generated by the SI mobility model will be reviewed first, followed by the other models. Given that the SI model had a targeted male student population of roughly 130 students, neither optimisation functions proposed an increase in supply. Despite the requirement for a few more seats, the optimisation goal has not altered because the smallest school type will enhance supply by 540 seats. This increase will shift the optimisation goal away from minimising the difference between the overall average of students and the supply level because the optimisation target is only 150 seats away from being optimum under the co-status. On the other hand, adding the smallest secondary school with 450 seats affects the optimisation aim by around 350 seats from its ideal state. However, the total targeted average for students for all other male projections is considerably closer to the minimal supply of 540 seats, so both optimisation functions raised the supply proportionately by adding one school of Type (18.5). See Table 35 for the schools' prototypes. A school building of type 18.5 has a building footprint area of 1003m2 and exclusively serves secondary stage education with 18 classrooms. It is worth noting that the parcel-specific optimisation algorithm (v1) employed two of the seven available parcels, despite the fact that this building type could be built on any land parcel due to its small footprint area. The present supply level for female students is above the ideal demand level in all situations. As a result, no attempt was made by the optimisation functions to boost the supply side. See Figure 46 for the results of District 124 secondary stage supply optimisation.

According to the ADA's 2030 action plan, the planned number of additional schools is two for boys and one for girls, with a total size of 36,500 m^2. When compared to ADA, our optimisation results may be considered extremely conservative yet correct in balancing the demand and supply of school seats.

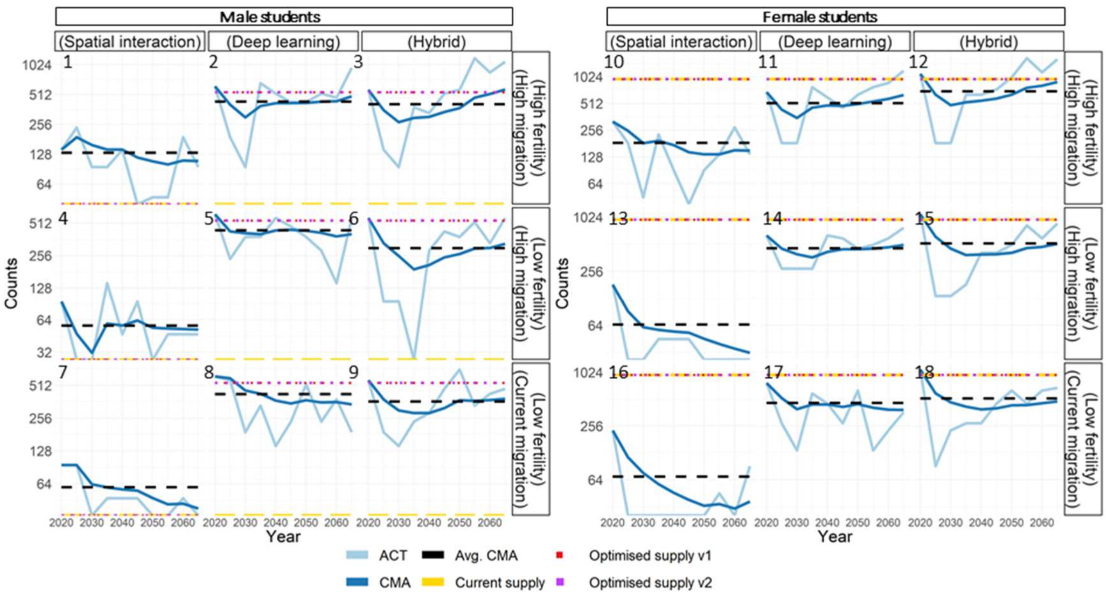


Figure 46 Results of District 124 secondary stage supply optimisation

District 124 has a reasonably constant population forecast, but it may become congested in the far future. Furthermore, the district still has the ability to accommodate new schools, but our estimates do not reveal an urgent need to provide for female secondary pupils. Given that all white land parcels for education are acceptable for hosting the chosen school type 18, the actual site must be determined based on factors other than land area, such as street width or proximity to other female schools in surrounding districts.

5.2.3.3 District 139

The entire area of District 139 is 244 km^2. In reality, the so-called district marks the limits of the Namar municipality, which is divided into six non-typical sub-districts. The aggregation of Namar districts in this model was achieved by the data curation described in section Built-Environment related data 3.8.2. As a result, District 139 does not reflect a conventional Riyadh district planning unit and may be dealt with at the municipal level. Furthermore, District 139 contains 59 white land parcels dedicated to education with a total area of 409696 m^2, an average parcel area of 11573 m^2, and a standard deviation of 2044 m^2. Furthermore, most estimates show a dramatically expanding population trend in this area, with the exception of the NN mobility model, which shows a falling population trend. Given the vast amount of available residential white land in this district and the expected rapid population growth in Riyadh, it is nearly impossible for this municipality-like district to experience population decline in the near or long term. As a result, demand for schools may be extraordinarily high in situations where population growth is projected to be significant. The estimated number of additional schools required using the non-parcel specific optimisation function ranges from 1103 to 1604 for IS, 4 to 14 for NN, and 527 to 1014 for hybrid residential mobility models. Despite the fact that the total number of schools added under the various models could be extremely high, district 139 has the potential to equal nearly 44% of the current centrally developed municipalities areas. Regardless of the actual pattern of population change in the area, the significant variation in targeted total student averages across the various scenarios provided a more accurate picture of the optimisers' behaviour. Before delving deeper into the specifics of each educational stage in district 139, it is critical to keep in mind the possibility of school buildings serving multiple education stages. The breakdown of schools' counts in the following sub-sections will count schools multiple times based on the number of education stages served. The following are the optimisation outcomes for each stage:

A. Elementary stage supply optimisation for district 139

The current elementary supply level for male students is 11076 whereas for females the supply level is 10357 seats. In most cases, the current supply level is expected to be below needed capacity except for projections using the NN mobility model where the current supply level could be sufficient. Given the unremarkable difference in the optimisation outcomes between male and female students, the results will be discussed based on residential mobility models and derivatives of population change. For the Both SI and hybrid mobility models, the numbers of students are expected to grow rapidly in the first 10 years before the growth slows for about 20 years and then accelerate again. See Figure 47 for results of District 139 elementary stage supply optimisation.

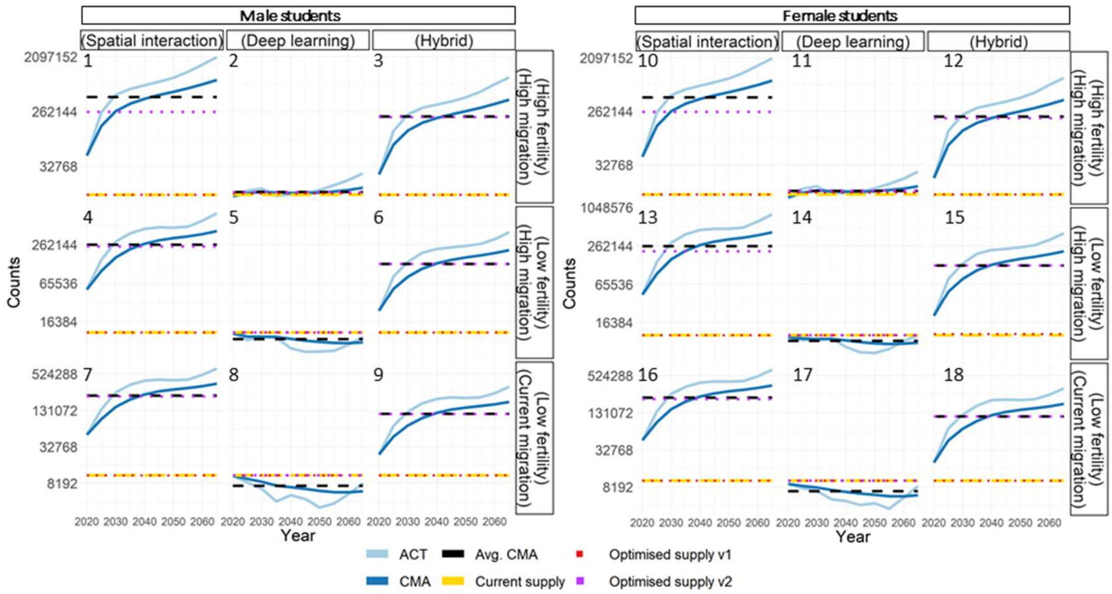


Figure 47 Results of District 139 elementary stage supply optimisation.

For the cases 1 and 10 (SI mobility with high fertility and high migration rates), the number of students is expected to grow from around 50 thousand to around 2 million, whereas the targeted average for supply is around 452000 seats for each gender. The parcel-specific optimisation function (v1) was tested for this task despite its known limit to provide building types for 10 different land parcels. However, the optimisation outcomes did not alter the supply for elementary education. The author hypothesises that the reason for not chaining the supply level is that the total percentage change was rounded to zero due to its small size. On the other hand, the non-parcel specific optimisation function (v2) increased the number of seats available to each gender by 246500. The optimisation function achieved 57% of the target capacity. Additionally, this scenario for each gender adds 150 dedicated elementary schools, 400 elementary and middle stage schools, and 50 all-stage schools, totalling 600 elementary schools. In this case, the optimisation function reached its maximum supply of 50 buildings for all types of schools, which includes elementary education.

Moving into cases 4 and 13 (SI mobility with low fertility and high migration rates), the actual numbers of students are expected to reach 800,000 students by 2070, whereas targeted average capacity is just above 260000 seats. The parcel specific optimisation function did not change the supply. However, the non-parcel specific model was able to supply additional 234215 male seats and 205580 female seats. Consequentially, the provision level achieves an excellent 94% and 83% of the targeted supply levels for male and female students respectively. Accordingly, this scenario adds 150 dedicated elementary schools, 343 elementary and middle stage schools, and 50 all-stage schools for a total of 543 elementary schools for male students. For female students, this scenario adds 150 dedicated elementary schools, 251 elementary and middle school, and 50 all-stage schools, bringing the total to 451 elementary schools.

The actual number of students by 2070 is expected to be about 65000 under scenarios 7 and 16 (SI mobility with low fertility and migration rates). The targeted average seating capacity is a little less than 230000 seats. Like the two earlier scenarios, the parcel specific optimisation function (v1) did not change the current supply level. The non-parcel specific variant, on the other hand, was able to provide an additional 212360 male seats and 208250 female seats. As a consequence, for male and female students, the provision level reaches 97 % and 95 % of the targeted supply levels, respectively. In addition, this scenario adds 150 dedicated elementary schools, 287 elementary and middle stage schools, and 11 all-stage schools, bringing the total number of elementary schools for male students to 448. This scenario adds 150 dedicated elementary schools, 300 elementary and middle schools, and 50 all-stage schools for female students, bringing the total number of elementary schools to 500.

For the scenarios generated using the NN mobility model, despite how unlikely it is to have a declining population projection in that area of the city, this example provides a useful insight on how the optimisation functions would behave under different circumstances to achieve the targeted supply levels. For cases 2 and 11 (NN mobility with high fertility and migration rates), the upper limits of students are projected to be a little above 25000 with a targeted average of around 12100 students based on the cumulative moving averaging of students counts. The parcel specific optimisation (v1) selected 8 different combinations of schools which yielded in total 14 schools. Under this solution four land parcels would have complex types of schools with different education stages. Moreover, the number of additional schools is divided equally between male and female but the choice of school types is the main factor for final capacities. Nevertheless, this solution narrows the supply gap to 5% for male students and 6% for female students. However, using the non-parcel specific optimisation function resulted in the addition of 14 schools, divided equally between both genders. Moreover, given the flexibility of choice in this model, the achieved supply targets for male and female students were at 100% and 101% respectively. Nevertheless, for the cases 5, 8, 14 and 17, the current supply limits are above the targeted cumulative average of students. Therefore, both optimisation results did not provide any additional capacities.

For projections using the hybrid mobility model, similarities in trends to the results of SI models can be noticed but with lower magnitude of change in students counts. For cases 3 and 12 (Hybrid mobility with high fertility and migration levels) while the highest number of students is around 960,000, the targeted cumulative average is around 215000 students. For the parcel specific optimisation (v1), the supply level remained with no change. However, for the non-parcel specific optimisation function (v2), the number of additional seats for male and female students were 199695 and 190730 respectively. The optimisation result covers the targeted supply level of male students by 97% and female students by 94%. In addition, this scenario adds 150 dedicated elementary schools, 291 elementary and middle stage schools, and 0 all-stage schools, bringing the total number of elementary schools for male students to 441. This scenario adds 150 dedicated elementary schools, 224 elementary and middle schools, and 50 all-stage schools for female students, bringing the total number of elementary schools to 424. Moreover, for all remaining cases (6, 9, 15 and 18), the parcel specific optimisation function (v1) only changed the supply of male students in case 6 due to selecting the school building type (121) which added a negligible 450 seats resulting in a total of 8% in supply level. On the other hand, the non-parcel specific optimisation function (v2) was able to match the targeted supply level completely at 100%. The total number of schools for cases 6, 9, 15 and 18 were 245, 229, 180 and 215 respectively.

When compared to the ADA action plan, our estimates appear imprudent. This is because the demand for schools up to the year 2030 is targeted to increase by 51 elementary schools, 20 for males and 30 for females. With a total area of 410,325 square metres. However, these numbers are subject to the population growth model used and not the optimisation function performance.

B. Middle stage supply optimisation for district 139

The current middle stage supply level for male students is 4004 seats whereas for females the supply level is at 4939 seats. In most cases, the current supply level is expected to be below needed capacity except for projections using the NN mobility model where the current supply level could be sufficient, given the slight difference in the optimisation outcomes between male and female students. For Both SI and hybrid mobility models, the numbers of students are expected to jump from the year 2025, then it would slow for the remaining projection time horizon. See Figure 48 for results of District 139 middle stage supply optimisation.

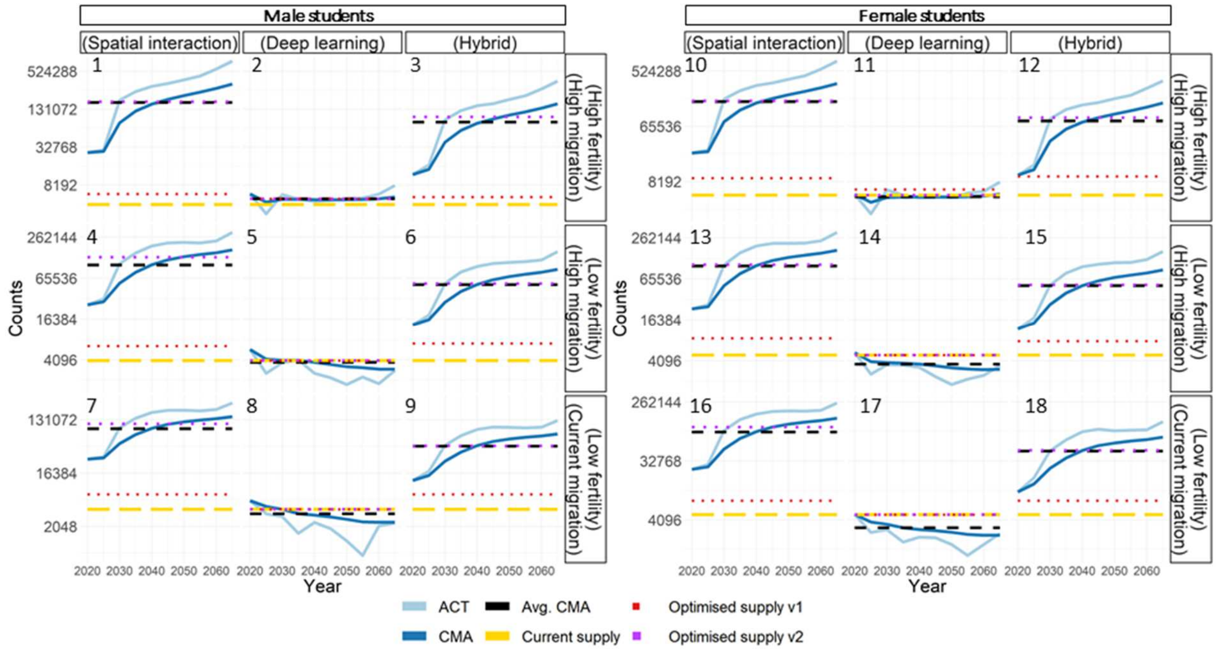


Figure 48 Results of District 139 middle stage supply optimisation.

For the projection using SI mobility modelling, the parcel specific optimisation did not perform well and its supply level averaged 7% with a standard deviation of 2% in all SI mobility modelling cases. On the other side, the non-parcel specific optimisation function surpassed the targeted average of supply in various degrees for all causes. While the male supply for high fertility with high migration, low fertility with high migration and low fertility with current migration were oversupplied by 3%, 30% and 21% respectively. The female oversupply was slightly better at 3%, 4% and 19%. Moreover, the total number of additional middle stage schools for both genders were 999, 684 and 648 respectively. To illustrate the size of oversupply, the 3% is equivalent to around 4500 seats and the 30% is about 31000 seats.

For the NN scenarios only instances 2 and 11 (high fertility with a high migration rate) had modifications to their current supply level, utilising the parcel-based optimisation function (v1). For male students, the supply level was 100%, while for female students, it was 132%. Prior to the optimisation, female oversupply stood at 7% and was increased by 1249 seats. However, results were improved when the non-parcel specific optimisation function (v2) was used. Male students received 100% of the intended supply and the female supply remained constant at 107%. To elaborate on the findings meanwhile, the parcel-specific optimisation function adds two school buildings, one dedicated to middle school students of type 15.5 and another to elementary and middle school students of type 33.11, totalling 945 seats. Female results include the utilisation of two middle stage dedicated schools of types 18.2 and 21.5, with a combined capacity of 1170 seats. On the other hand, the non-parcel specific modification includes one middle stage dedicated school of type 18.2 and two elementary schools with middle stage schools of types 540.18 and 9.24, for a total of 945 male seats, which is consistent with the parcel specific outcomes. For female students, no middle stage schools were introduced.

Regarding the Hybrid mobility modelling outcomes. the parcel specific optimisation (v1) did not perform well, and its supply level averaged 13% with a standard deviation of 3% in all hybrid mobility modelling cases. On the other side, the non-parcel specific optimisation function surpassed the targeted average of supply in all causes, while the male supply for high fertility with high migration, low fertility with high migration and low fertility with current migration were oversupplied by 21%, 3% and 1% respectively. The female oversupply was slightly better at 13%, 3% and 2%. Additionally, the total number of additional middle stage schools for both genders were 565, 282 and 213 respectively. To illustrate the size of oversupply, 3% is equivalent to around 1400 seats and 21% is about 17000 seats.

When compared to the ADA action plan, our projections appear to be excessive. This is because the need for schools is expected to expand by 40 middle schools, 21 for boys and 19 for girls, between now and 2030. With a total area of 423,600 m2.

C. Secondary stage supply optimisation for district 139

The current secondary stage supply level for male students is 1843 seats whereas the supply level for females is 1766 seats. Moreover, in all cases, the current supply level is expected to be below the needed capacity which highlights the need for optimisation. The numbers of both elementary and middle stages students are expected to jump starting from the year 2030. See Figure 49 for the results of District 139 secondary stage supply optimisation.

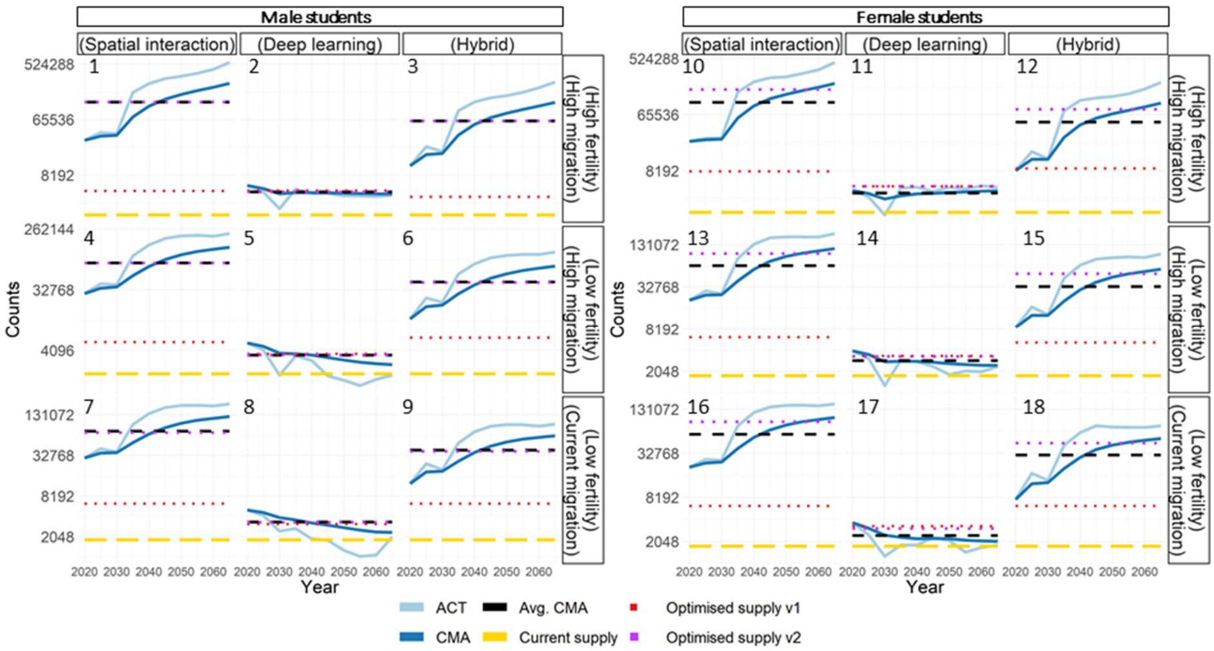


Figure 49 Results of District 139 secondary stage supply optimisation.

For the male students' examples (1, 4, and 7) created with the SI mobility model and for the parcel specific model (v1), given the required size of supply, the ten parcels solution provided 10% of the needed seats at its best with an average of 8% for all male and female cases. Moreover, the school type 165 was chosen for all parcels but with different ratios of male and female combinations. Overall, the parcel specific (v1) outcomes should not be used for this level of high demand. On the other hand, the non-parcel specific (v2) optimisation findings for high fertility with high migration, low fertility with high migration, and low fertility with current migration met their targets by 100%, 100%, and 95%, respectively. The 5% supply deficit is equivalent to 3700 seats. However, the optimisation for female students in instances (10, 13 and 16) shows a significant oversupply of 61%, 49%, and 48%, respectively. The number of female seats that are oversupplied ranges from over 28883 to 62540. Although these levels of oversupply appear to be unacceptable, they are within the maximum range of seats needed for both actual projection numbers and their cumulative move averages for each five-year period. In reality, comparing the average of actual forecasts to the optimisation findings would result in an undersupply of 21%, 17%, and 14% respectively. To address this issue, increasing the number of iterations is likely to yield better outcomes.

Moving into the NN outcomes, on the one hand, for male students in cases (2, 5 and 8), the parcel specific optimisation (v1) averaged 101% with a standard deviation of 5%, whereas the non-parcel specific (v2) averaged 100% with a standard deviation of 1%. On the other hand, for female students in cases (11, 14 and 17) while the parcel specific solution (v1) averaged 127% with a standard deviation of 7%, the non-parcel specific solution (v2) averaged 123% with a standard deviation of 6%. To have a detailed understanding of one of these optimisation outcomes, see Table 53 for the optimisation outcomes for the secondary stage projections using the NN model with high fertility and migration rates.

Table 53 the optimisation outcomes for the secondary stage projections using the NN model with high fertility and migration rates.

Gender	Opti. function	ID of school type	Count	Secondary stage capacity	Attached educational buildings	Secondary Building footprint area (m^2)	m^2/student
Male	V1	150:(26.2)	1	780	em - s	1862	2.39
		98:(20.62)	1	600	e - m - s	1797	3
		182:(26.26)	1	780	s	2072	2.66
		183:(20.26)	1	600	s	1797	3
	V2	24.05	1	720	s	1003	1.39
		30.11	2	900	s	5472	6.08
Female	V1	179:(24.5)	1	720	s	1003	1.39
		31:(18.5)	1	540	e - m - s	1003	1.86
		184:(30.11)	1	900	s	5472	6.08
		160:(24.5)	1	720	m - s	1003	1.39
	V2	18.05	4	540	s	1003	1.86
		24.05	1	720	s	1003	1.39

V1 parcel specific optimisation
V2 non-parcel specific optimisation
s secondary stage in one building
em elementary and middle schools in one building
e - m – s elementary middle and secondary schools in separate buildings
m middle stage schools in one building
* All numbers are rounded

Table 53 demonstrates how dissimilar the optimisation outputs are despite their proximity to the defined objective. The parcel specific optimisation function (v1) supplied an additional 2760 seats for male students by assigning four school buildings to three building types (26.2, 20.62, and 20.26), requiring a total footprint area of 7528 m^2. In comparison, the non-parcel specific optimisation (v2) function adds 2520 seats from three school buildings of types (24.05, 30.11), with a total footprint area of 11947 m2. Both (v1) and (v2) optimisation functions produced excellent results, delivering 105% and 100% of the required seats, respectively. For female supply, the (v1) optimisation function added 2880 seats by adding four schools with three distinct school types (24.05, 18.05, and 30.11), requiring a footprint area of 8481 m^2. Whereas the (v2) optimisation function provided 2880 seats by adding 5 buildings with only two school types (18.05

and 24.05), which required a total footprint area of 5015 m^2. In this example, both female optimisation functions exceeded the desired capabilities by 29%. It should be emphasised, however, that the parcel-specific restriction was not based on the size of the land. Instead, it was removed to find the best solution. The rationale for this is that the average area of all 59 available lands is 6899 m^2, which is about double the average area of all 184 schooling combinations in optimisation v1. As a result, the probability of discovering an infeasible solution is very low, and the optimisation constraint for land-specific solutions has been eased.

The hybrid mobility model was used to produce the projections for male students (3, 6, and 9). The non-parcel specific optimisation (v2) findings for high fertility with high migration, low fertility with high migration, and low fertility with current migration all achieved their aims by 100%, 98%, and 96 %, respectively. The supply gap of 3.76% equates to 1472 seats. However, optimisation for female students in instances (12, 15, and 18) reveals a large oversupply of 61 %, 51 %, and 45 %, respectively. The number of oversupplied female seats ranges from over 14107 to 30446. Although these levels of oversupply appear to be problematic, they are within the maximum range of seats required by both the actual projected numbers and their cumulative moving averages across the projection time horizon. Comparing the average of actual prediction values to the optimisation findings would result in a 50%, 42.77%, and 41.16 % undersupply, respectively. Given the required size of supply for the parcel specific model (v1), the ten-parcel solution offered 20 % at maximum of the needed seats, with an average of 15 % for all male and female cases. Furthermore, school type 165 was dominant throughout all parcels, but with varying male/female ratios. Overall, the parcel-specific approach is unsuitable for such large demand.

The demand for additional schools by the ADA action plan is expected to be 27 secondary schools by 2030, 14 for male and 13 for female students with a total area of 330,500m2.

5.2.4 Remarks on the District Level Optimisation

The earlier optimisation outcomes might be suggesting a better performance for the non-parcel specific optimisation function (v2) over the parcel specific optimisation function (v1), given its overall better achieved level of supply. However, the results are subject to the demand size. On the one hand, where the number of surplus students is beyond the maximum capacity of the parcel specific optimisation function the results of the non-parcel specific optimisation (v2) are better and justify its development. On the other hand, when the size of demand is within the limitation of the parcel specific function (v1) the gap in the performance is small and the models outcomes could be identical as shown in optimisation for district 124 for the female secondary stage supply using SI mobility.

Given the quality of optimisation, picking which optimisation to use is challenging. This is due to the fact that other parameters such as total built-up area, development expenses, proximity to existing residential area and total number of schools must also be considered. In other words, given the existing optimisation configuration, which takes into account all three schooling phases for both genders at the same time, the generalised level of optimisation becomes more important than stage-specific or gender-specific optimisations. In theory, increasing the number of iterations of the optimisation function should improve the likelihood of discovering a better solution. Separating each optimisation task should also enhance model performance, but the complex consideration of shared land areas by multiple schools will be more difficult to accomplish. Furthermore, the time and setup costs will be greater for the final user.

Chapter 6

Conclusion and Future Work

Despite the apparent flaws in present urban planning practice, the field's advancement appears to have reached a deadlock for an extended period of time, particularly when it comes to the criteria for the supply of public amenities and commercial activity. The problems of traffic congestion, urban sprawl, and pollution in today's cities are well-known, but the solutions are typically ad hoc and based on conceived sustainable ideas like increased green space and clean public transportation, or in some cases, where land use transportation models are applied, the primary focus is on transportation and environmental issues. But another issue that is typically disregarded in modern cities is the uneven supply of public amenities and commercial services. Today's urban planners appear to be less interested in long-term strategic planning and more concerned with creating visually appealing master plans that follow global design norms. To further explain this matter, several roles exist for urban planners and urban designers during the planning process. Long-term strategic plans that direct the expansion and growth of cities and regions must be created by urban planners. To determine the requirements of the community and create plans that address those needs, they collaborate with local governments, community organisations, and other stakeholders. On the other hand, the built environment's physical design is the responsibility of urban designers. They strive to design environments that are both aesthetically pleasing and practical for the community. These fields complement one another since they both emphasise on adaptability, sustainability and safety of cities. Yet, in order to prevent the current overlap in practise, the objectives of various disciplines must be clearly defined. In the best-case scenario, which is thought to be uncommon, interaction modelling of some kind is examined, but for reasons other than balancing the demand and supply of facilities and services. By taking the usage of PSS to a new level, land use guidelines could be developed more effectively. However, creating these kinds of support systems is a difficult endeavour, mainly because there is a lack of data and empirical research on the main elements affecting the sub processes. In addition to these two basic problems, there may be a number of other problems that make it difficult to build and run PSS at a disaggregated level. It is possible that the perceived difficulty in developing the necessary planning support tools is due to the poor quality of the data that is currently available, the difficulty in persuading sceptics to change their minds, and the time and effort needed to involve numerous subject matter experts**.** This is thought to discourage urban planners from using the modelling approach. There is enough information to build the tools, and PSS can be used to solve or mediate many urban problems, thus the effort in this area shouldn't ever be given up. Since it is doubtful that there will ever be a perfect solution for every issue, near-optimal solutions are usually preferred as a development strategy. Furthermore, the employed measurements or standards must be re-examined, enhanced, or even altered to more truly reflect the concerns being addressed. In order to revolutionise the area of urban planning, urban planners must be innovative once more but this time with greater sophistication.

The PSS tool developed for this study was meant to be as generic and modular as possible. The concept of generalisability encompasses the application to other cities and communities, as well as the investigation of additional issues outside the provision of facilities criteria. This feature is achieved by incorporating what are believed to be the urban system's most important sub-models, such as population growth, residential mobility, and land development. In a sense, it is easier to plan for the demands of the population once the population and their movement in the future are known. Additionally, by fully incorporating several sub-models, the correlation and causation of some elements among the many models may be better understood, serving the model's explanatory side in addition to its forecasting purpose. On the other hand, the model's modularity was applied by developing distinct programming classes for each aspect, such as Household containers, separate family members, and Districts. Then, unique functions can be added or modified as needed within each class. However, it must be acknowledged that the data restrictions were substantial and significantly influenced the model's current form. For additional developers to contribute to the development of this model, it is necessary for them to possess a minimum level of programming expertise.

Numerous lessons were learned during the course of this research, which are expected to be applicable to any future urban modelling endeavours. First, the importance of data quality exceeds that of the type of model. Prior to attempting to model any portion of an urban system, it is vital to examine the data availability, collection frequency, intrinsic consistency, predicted error levels, level of representation of the desired behaviour or trend, ability to make synthesis data, and data permissions. This is due to the fact that the outcomes of the established models rely heavily on the quality of the inputs and are not intended to fix the data. Second, data forecasts based on historical patterns, such as fertility factors or land prices, cannot be evaluated only on the basis of goodness of fit, as this metric could be deceptive. Indeed, the historical depth for trend extrapolation in this study was often limited. But the unanticipated outcomes of the automated projection methods, which are based on error measures, became problematic, with land prices dropping to zero and other implausible shifts. For greater precision, a case-by-case analysis of these estimates is required. Third, for each administratively-defined geographical boundary, the local differences within each boundary must be analysed. This is due to the fact that distinct localities/characteristics can be found within each location, with varying levels of appeal among the diverse population groups. Local disparities in land prices, for instance, might indirectly affect a variety of interactions involving the location of people and business. In some circumstances, a redefinition of the geographical boundaries may be necessary after a thorough analysis, or an optimal planning unit size may be identified to significantly reduce the disparities. Fourth, the cooperation of topic experts is thought to be necessary for the modelling of urban systems. This includes urban planners, mathematicians, programmers, demographers, real estate developers, law makers, decision makers, and a variety of other professionals. The volume and variety of judgements that must be based on higher levels of knowledge necessitated the division and distribution of work across a team with the objective of resolving the complex issues in each aspect. For instance, demographers would examine the derivatives of population increase and predict the most probable growth scenarios. Equally, programmers could modify the design of models to make them more modular, integrable, computationally efficient, and quick to model. Fifth, in addition to providing a visual representation of the model's interactions and results, visualising every facet of the model is essential for identifying modelling flaws. Due to the vast amount of numbers generated by the various models, it may be impossible to identify all of the inaccuracies in tabular form. For instance, after being plotted, the implausible trends in fertility predictions were quickly identified. In addition, while developing the ABM, some problems may be generated by coding and others by backend memory management. In both instances, it is difficult to detect these problems without a visualisation tool. During modelling of household mobility, it was discovered that household members resided in different districts than their household containers. This was only detectable by visualising the connections between the household agents as physical lines. This issue was caused by the fact that AnyLogic's internal memory management needed to separate the movement events of household containers and members. Sixth, the used decision functions at the individual level are of utmost importance, and despite the level of success that could be achieved by employing Logit models, Neural Networks, and Gravity models, other models such as Analytical Hierarchical Processing should be tested for potential advantages such as interpretability of parameters, ease of calibration, and increased prediction accuracy.

The findings of the study and the policy implications for both researchers and decision-makers can be summed up as follows. First, the research findings at this point are indicative and suggestive of the model's performance rather than necessarily providing a precise picture of the effects of policy changes. Second, because the projection confidence of such models typically decreases over time, constant evaluation and calibration are part of the optimisation process for practitioners. The long-term predictions are useful for scenario testing, such as the assessment of best and worst case scenarios, the study suggests, although the projection of the near future, which covers a period of 10-15 years, may have useful results if valid data is used. Third, for the optimisation process to be successful, the type of standard and its vital components must be defined and unified. This needs to be done at all levels with the various connected entities providing (like MOE) or evaluating the service (i.e. ADA). This is due to the fact that comparing several standards does not produce relevant feedback and makes it challenging to communicate amongst the various bodies. Forth, similar to other decision-making processes, selecting the appropriate optimization function depends on the factors taken into account and the objectives established by the decision-makers to evaluate the service. Not every function is applicable in every situation. Fifth, to progress these kinds of models in terms of its explanatory or forecasting functions, researchers must first increase their understanding of the primary derivatives of population change and their interrelations.

When it comes to future development, many points might be focused upon. Firstly, the request for improved data quality must be made to the relevant entities. This is to call for standardised data that is gathered more regularly, is well-defined and consistent, has more information regarding residential mobility and internal immigration, offers more information about expected error levels, and is open at the district level. On the one hand, residential mobility should be thoroughly covered in data gathering in order to comprehend the impact of lifecycle changes on mobility triggers and also the economic considerations that determine the locations shortlisted for movement. In addition, distance thresholds to elements such as social life, place of employment, and shopping centres are considered while selecting a residence. Internal migration must also be covered to understand the ages, marital status, and family size of migrants, as well as any patterns of reversed movement at key ages. Secondly, the standards employed to provide the various facilities and the aims of the provision must be agreed upon by the relevant governing authorities. Whether the criteria are based on area per student or pre-designed buildings, the main components and upper and lower limits of the facility in question must be clearly defined and specified. By formalising the standards' calculation base and its operational limitations, the standards' values should become more relevant and applicable. Thirdly, a more effective population synthesiser should be created to represent the intended population. Both the sampled base-year population and the migrated population should be synthesised with greater precision and specificity. For instance, the change-by-replacement strategy for households at the district level does not take into consideration additional characteristics outside age structure and gender. Aspects such as income, previous residence, and education, which are essential for profiling population behaviour, are ignored. In addition, despite the paucity of data about the migrated population, it is anticipated that more information may become accessible in the future; hence, a household synthesiser will be necessary to create these extra households and appropriately include them. The current model comprises solely the Saudi population on the grounds that they constitute a closed population system. However, the requirements of the non-Saudi people must be incorporated into the model. This inclusion may be implemented by calibrating a parallel model with only the non-Saudi population, or the current model could be modified by incorporating the non-Saudi population using identifying parameters and distinct behavioural functions. Fourthly, the existing optimising functions only take school additions into account. The deduction by demolition or use change is an additional factor to consider. In addition, the optimisation of a single land use could be expanded to cover numerous land uses that consume a shared land resource. These modifications will make this procedure more comprehensive and closer to the ideal of land use planning. Fifthly, the modularity of the model can be enhanced by employing a shared data store in a manner similar to that of UrbanSim. Currently, the created model's data is accessed from within the individual programming classes, but ideally, a shared store would facilitate easier and more efficient information interchange between the existing sub-models and any new ones. Sixthly, despite the fact that the residential mobility triggers are tuned to current trends, there is a strong perception that they are occurring faster than is actually the case. Therefore, with the aid of more data, the residential mobility triggers should be investigated further and confined to recognised limitations in order to achieve more accurate modelling results.

Reference list

Abdulaal, W.A. (2012) 'Large urban developments as the new driver for land development in Jeddah', *Habitat International*, 36(1), pp. 36–46. doi: 10.1016/j.habitatint.2011.05.004

Abel, G.J. and Sander, N. (2014) 'Quantifying global international migration flows', *Science (New York, N.Y.)*, 343(6178), pp. 1520–1522. doi: 10.1126/science.1248676

Acheampong, R.A. and Silva, E. (2015) 'Land use–transport interaction modeling: A review of the literature and future research directions', *Journal of Transport and Land Use*, 8(3), pp. 11–38. doi: 10.5198/jtlu.2015.806

ADA (2011) *The Action Plan for Coordination and Provision of Public Services in Riyadh (Executive Summary) 2011*. Riyadh.

ADA (2015) *ArRiyadh Development Authority: Planning - The Metropolitan Development Strategy for Arriyadh Region* (Accessed: 29 November 2017).

Aggarwal, C.C. (2018) *Neural Networks and Deep Learning*. Cham: Springer International Publishing.

Akimoto, F. (2009) 'The birth of 'land use planning' in American urban planning', *Planning Perspectives*, 24(4), pp. 457–483. doi: 10.1080/02665430903145705

Al Naim, M.A. (2013) 'Urban Transformation in the City of Riyadh: A Study of Plural Urban Identity', *Open House International*, 38(4), pp. 70–79. doi: 10.1108/OHI-04-2013-B0008

Al-Hathloul, S. (2017) 'Riyadh Development Plans in the Past Fifty Years (1967-2016)', *Current Urban Studies*, 05(01), pp. 97–120. doi: 10.4236/cus.2017.51007

Al-Olet, A.A. (2004) 'The development process of the approval methods of land subdivision plans in Riyadh city', *Journal of King Saud University: Architecture & Planning*, 17, 1--50. Available at: http://cap.ksu.edu.sa/sites/cap.ksu.edu.sa/files/imce_images/17_1e_0.pdf.

Alonso, W. (1960) 'A THEORY OF THE URBAN LAND MARKET', *Papers in Regional Science*, 6(1), pp. 149–157. doi: 10.1111/j.1435-5597.1960.tb01710.x

AlSaif, S. (2016) *Assessing the current level of Service Model: Automated Tool Methodology (Heat map)*.

Alskait, K. (2003) ‘SUBDIVISION PLANNING IN RIYADH: PROBLEMS & REMIDIES’, *Emirates Journal for Engineering Research*, 8(2), 39--50.

Anas, A. (1983) ‘Discrete choice theory, information theory and the multinomial logit and gravity models’, *Transportation Research Part B: Methodological*, 17(1), pp. 13–23. doi: 10.1016/0191-2615(83)90023-1

Andridge, R.R. and Little, R.J.A. (2010) ‘A Review of Hot Deck Imputation for Survey Non-response’, *International Statistical Review*, 78(1), pp. 40–64. doi: 10.1111/j.1751-5823.2010.00103.x

April, J. *et al.* (2003) ‘Practical introduction to simulation optimization’, *Proceedings of the 2003 Winter Simulation Conference.* New Orleans: Citeseer, pp. 71–78. doi: 10.1109/WSC.2003.1261410

APS (2021) *School Design Standards*. Available at: https://www.aps.edu/facilities-design-and-construction/documents/design-standards-and-guidelines/HS_Standards.pdf (Accessed: 15 July 2022).

Araújo de Oliveira, V.M. (2022) *Urban Morphology*. Cham: Springer International Publishing.

Arentze, T., Timmermans, H. and Veldhuisen, J. (2010) ‘The Residential Choice Module in the Albatross and Ramblas Model Systems’, in Pagliara, F., Preston, J. and Simmonds, D. (eds.) *Residential Location Choice.* (Advances in Spatial Science). Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 209–222.

Auchincloss, A.H. and Diez Roux, A.V. (2008) ‘A new tool for epidemiology: the usefulness of dynamic-agent models in understanding place effects on health’, *American Journal of Epidemiology*, 168(1), pp. 1–8. doi: 10.1093/aje/kwn118

Axelrod, R. (1997) ‘Advancing the Art of Simulation in the Social Sciences’, in Conte, R., Hegselmann, R. and Terna, P. (eds.) *Simulating Social Phenomena.* (Lecture Notes in Economics and Mathematical Systems, 456). Berlin, Heidelberg: Springer Berlin Heidelberg; Imprint: Springer, pp. 21–40.

Ballas, D., Broomhead, T. and Jones, P.M. (2019) ‘Spatial Microsimulation and Agent-Based Modelling’, in Briassoulis, H., Kavroudakis, D. and Soulakellis, N. (eds.) *The Practice of Spatial Analysis.* Cham: Springer International Publishing, pp. 69–84.

Ballas, D. and Clarke, G.P. (2001) 'Modelling the Local Impacts of National Social Policies: A Spatial Microsimulation Approach', *Environment and Planning C: Government and Policy*, 19(4), pp. 587–606. doi: 10.1068/c0003

Ballas, D., Rossiter, D., Thomas, B., Clarke, G. and Dorling, D. (2005) *Geography matters: Simulating the local impacts of national social policies*. York: Joseph Rowntree Foundation.

Barnard, P.O. (1986) *Modelling shopping destination choices : a theoretical and empirical investigation*. (Doctoral dissertation).

Bartholomew, H. (1932) *Urban Land Uses: Amounts of Land Used and Needed for Various Purposes by Typical American Cities. An Aid to Scientific Zoning Practice*. 2014th edn. (Harvard City Planning Studies, 4). Cambridge, MA: Harvard University Press.

Batty, M. (2007) 'Model cities', *The Town Planning Review*, 78(2), 125-151. Available at: http://www.complexcity.info/files/2011/06/batty-tpr-2007.pdf (Accessed: 18 June 2022).

Batty, M. (2012) 'A Generic Framework for Computational Spatial Modelling', in Heppenstall, A.J. *et al.* (eds.) *Agent-Based Models of Geographical Systems.* Dordrecht: Springer Netherlands, pp. 19–50.

Ben-Akiva, M. and Bierlaire, M. (2012) 'Discrete Choice Models with Applications to Departure Time and Route Choice', in Hall, R. (ed.) *Handbook of transportation science.* (23): Springer Science & Business Media, pp. 7–37.

Ben-Akiva, M.E. and Lerman, S.R. (1985) *Discrete choice analysis: Theory and application to travel demand*. (MIT Press series in transportation studies, 9). Cambridge Mass.: MIT Press.

Ben-Dor, G., Ben-Elia, E. and Benenson, I. (2021) 'Population downscaling in multi-agent transportation simulations: A review and case study', *Simulation Modelling Practice and Theory*, 108, p. 102233. doi: 10.1016/j.simpat.2020.102233

Birkin, M. and Clarke, M. (1988) 'Synthesis—A Synthetic Spatial Information System for Urban and Regional Analysis: Methods and Examples', *Environment and Planning A: Economy and Space*, 20(12), pp. 1645–1671. doi: 10.1068/a201645

Birkin, M. and Wu, B. (2012) 'A Review of Microsimulation and Hybrid Agent-Based Approaches', in Heppenstall, A.J. *et al.* (eds.) *Agent-Based Models of Geographical Systems.* Dordrecht: Springer Netherlands, pp. 51–68.

Boehm, T.P. (1982) 'A Hierarchical Model of Housing Choice', *Urban studies*, 19(1), pp. 17–31. doi: 10.1080/00420988220080021

Bohk-Ewald, C., Li, P. and Myrskylä, M. (2018) 'Forecast accuracy hardly improves with method complexity when completing cohort fertility', *Proceedings of the National Academy of Sciences of the United States of America*, 115(37), pp. 9187–9192. doi: 10.1073/pnas.1722364115

Bongaarts, J. and Feeney, G. (1998) 'On the Quantum and Tempo of Fertility', *Population and Development Review*, 24(2), pp. 271–291. doi: 10.2307/2807974

Booth, H. (2006) 'Demographic forecasting: 1980 to 2005 in review', *International Journal of Forecasting*, 22(3), pp. 547–581. doi: 10.1016/j.ijforecast.2006.04.001

Booth, H., Maindonald, J. and Smith, L. (2002) 'Applying Lee-Carter under conditions of variable mortality decline', *Population Studies*, 56(3), pp. 325–336. doi: 10.1080/00324720215935

Borshchev, A. (2013) *The Big book of simulation modelling: Multimethod modeling with Anylogic 6* [Hampton, NJ]: AnyLogic North America.

Brasington, D.M. (2021) 'Housing Choice, Residential Mobility, and Hedonic Approaches', in Fischer, M.M. and Nijkamp, P. (eds.) *Handbook of Regional Science.* Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 449–466 (Accessed: 15 April 2022).

Briassoulis, H., Kavroudakis, D. and Soulakellis, N. (eds.) (2019) *The Practice of Spatial Analysis*. Cham: Springer International Publishing.

Bröcker, J. (2021) 'Spatial Interaction Models: A Broad Historical Perspective', in Fischer, M.M. and Nijkamp, P. (eds.) *Handbook of Regional Science.* Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 99–123. Available at: https://link.springer.com/content/pdf/10.1007%2F978-3-662-60723-7_131.pdf (Accessed: 15 April 2022).

Brownstone, D. (2001) *Discrete choice modeling for transportation*. Available at: https://escholarship.org/uc/item/29v7d1pk.

Butz, W.P. and Ward, M.P. (1979) 'Will US Fertility Remain Low? A New Economic Interpretation', *Population and Development Review*, 5(4), p. 663. doi: 10.2307/1971976

Cascetta, E. (2009) *Transportation systems analysis: Models and applications*. 2nd edn. (Springer optimization and its applications, 29). New York: Springer.

Castle, C. and Crooks, A. (2006) 'Principles and Concepts of Agent-Based Modelling for Developing Geospatial Simulations', *CASA Working Papers*, 110, pp. 1–60. Available at: https://discovery.ucl.ac.uk/id/eprint/3342/1/3342.pdf (Accessed: 23 June 2022).

CDE (2000) *Guide to School Site Analysis and Development,* 15 July. Available at: https://www.cde.ca.gov/ls/fa/sf/guideschoolsite.asp (Accessed: 15 July 2022).

Cenani, Ş. (2021) 'Emergence and complexity in agent-based modeling: Review of state-of-the-art research', *Journal of Computational Design*, 2(2), 1-24. doi: 10.53710/jcode.983476

Cerdá, M. *et al.* (2014) 'Addressing population health and health inequalities: the role of fundamental causes', *American Journal of Public Health*, 104 Suppl 4(S4), S609-619. doi: 10.2105/AJPH.2014.302055

Chen, L. (2012) 'Agent-based modeling in urban and architectural research: A brief literature review', *Frontiers of Architectural Research*, 1(2), pp. 166–177. doi: 10.1016/j.foar.2012.03.003

CIoP (2013) *The Canadian Institute of Planners: What is Planning?,* 11 July. Available at: https://web.archive.org/web/20131221215054/http://www.cip-icu.ca/web/la/en/pa/3FC2AFA9F72245C4B8D2E709990D58C3/template.asp (Accessed: 11 July 2022).

Clarke, R. (2016) 'Big data, big risks', *Info Systems J*, 26(1), pp. 77–90. doi: 10.1111/isj.12088

Cochrane, R.A. (1975) 'A Possible Economic Basis for the Gravity Model', *Journal of Transport Economics and Policy*, 9(1), pp. 34–49. Available at: http://www.jstor.org/stable/20052391.

Coleman, J.S. (1994) *Foundations of social theory*: Harvard University Press.

Coulombel, N. (2011) *Residential choice and household behavior: State of the Art*. Cachan, France (SustainCity Working Paper 2.2a).

Crooks, A.T. and Heppenstall, A.J. (2012) 'Introduction to Agent-Based Modelling', in Heppenstall, A.J. *et al.* (eds.) *Agent-Based Models of Geographical Systems.* Dordrecht: Springer Netherlands, pp. 85–105.

Crooks, A.T., Patel, A. and Wise, S. (2014) 'Multi-Agent Systems for Urban Planning', in Pinto, N. *et al.* (eds.) *Technologies for urban and spatial planning: Virtual cities and territories.* Hershey: Information Science Reference, pp. 29–56.

Daghistani, A.-M.I. (1985) *Ar-Riyadh: urban development and planning*: Kingdom of Saudia Arabia, Ministry of Information, Interior Information.

Darin-Drabkin, H. (1977) *Land Policy and Urban Growth*. (Urban and regional planning series, 16). Oxford: Elsevier.

DFE (2014) *Area guidelines for mainstream schools*. UK (Building Bulletin 103). Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/905692/BB103_Area_Guidelines_for_Mainstream_Schools.pdf (Accessed: 11/May/2021).

Dieleman, F.M. (2001) 'Modelling residential mobility; a review of recent trends in research', *Journal of Housing and the Built Environment*, 16(3), pp. 249–265. doi: 10.1023/A:1012515709292

Ding, S. *et al.* (2013) 'Evolutionary artificial neural networks: a review', *Artificial Intelligence Review*, 39(3), pp. 251–260. doi: 10.1007/s10462-011-9270-6

Duany, A., Sorlien, S. and Wright, W. (2009) 'SmartCode Version 9.2', *Ithaca: New Urban News Publications Inc*.

Dudek, M. (2007) *Schools and Kindergartens: A Design Manual*. Basel: Birkhäuser Basel.

El Mallakh, R. and El Mallakh, D.H. (1982) *Saudi Arabia: Energy, developmental planning, and industrialization*. Lexington, MA: Lexington Books.

Feldman, O. *et al.* (2010) 'A Microsimulation Model of Household Location', in Pagliara, F., Preston, J. and Simmonds, D. (eds.) *Residential Location Choice.* (Advances in Spatial Science). Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 223–241.

Fischer, M.M. and Aufhauser, E. (1988) 'Housing Choice in a Regulated Market: A Nested Multinomial Logit Analysis', *Geographical Analysis*, 20(1), pp. 47–69. doi: 10.1111/j.1538-4632.1988.tb00162.x

Fleten, S.-E., Høyland, K. and Wallace, S.W. (2002) 'The performance of stochastic dynamic and fixed mix portfolio models', *European Journal of Operational Research*, 140(1), pp. 37–49. doi: 10.1016/S0377-2217(01)00195-3

Fotheringham, A.S. *et al.* (2004) ‘The Development of a Migration Model for England and Wales: Overview and Modelling Out-Migration’, *Environment and Planning A: Economy and Space*, 36(9), pp. 1633–1672. doi: 10.1068/a36136

GAS (1992) *Some Highlights of Population Census 1413H (1992) compared to the 1394 Census*.

GAS (2001) *Population characteristics in the Kingdom of Saudi Arabia from the results of demographic research 2001)*. Available at: https://beta.stats.gov.sa/sites/default/files/2019-07/ar-demographic-research-1999.pdf.

GAS (2004) *Highlights Population & Housing Census 1425H (2004)*.

GAS (2007) *Highlights Demographic Survey 1428H (2007)*. Available at: https://beta.stats.gov.sa/sites/default/files/2019-07/Demographic2007.pdf.

GAS (2010) *General Census of Population and Housing 1431H - 2010: Detailed results of the Kingdom*. Available at: https://beta.stats.gov.sa/sites/default/files/2019-04/ar-census2010-dtl-result_2_1_0.pdf.

GAS (2016) *Demography Survey (2016)*. Available at: https://beta.stats.gov.sa/sites/default/files/2019-07/ar-demographic-research-2016_4.pdf.

GAS (2017) *The Sixteenth Services Guide 2017: Ar-Riyad Region*.

GAS (2020) *Marriage and Divorce Statistics 2020*. Available at: https://www.stats.gov.sa/sites/default/files/Marriage%20and%20Divorce%20Statistics%202020%20EN.pdf.

Gilbert, N. and Bankes, S. (2002) ‘Platforms and methods for agent-based modeling’, *Proceedings of the National Academy of Sciences*, 99(suppl_3), pp. 7197–7198. doi: 10.1073/pnas.072079499

Gilbert, N. and Doran, J. (eds.) (2018) *Simulating Societies*. London: Routledge.

Girosi, F. and King, G. (2007) ‘Understanding the Lee-Carter mortality forecasting method’, *Gking. Harvard. Edu*.

Glover, F., Kelly, J.P. and Laguna, M. (1996) ‘New advances and applications of combining simulation and optimization’, *Winter Simulation Conference*, pp. 144–152. doi: 10.1145/256562.256595

GOA (2012) *DESIGN MANUAL: Minimum Requirements for Private School Facilities*. A Guide for School Owners, Administrators and Design Professionals - November. 403rd edn. Abu Dhhabi (DESIGN MANUAL). Available at: https://jawdah.qcc.abudhabi.ae/en/Registration/QCCServices/Services/STD/ISGL/ISGL-LIST/PR-403.pdf.

Gorman, D.M. *et al.* (2006) 'Agent-based modeling of drinking behavior: a preliminary model and potential applications to theory and practice', *American Journal of Public Health*, 96(11), pp. 2055–2060. doi: 10.2105/AJPH.2005.063289

GOV.UK (no date) *Make a neighbourhood plan*. Available at: https://www.gov.uk/government/get-involved/take-part/make-a-neighbourhood-plan (Accessed: 13 April 2022).

GOV.UK (2017) *Guidance: Local Palns*. Available at: https://www.gov.uk/guidance/local-plans--2 (Accessed: 15 January 2018).

Grant, J. (2002) 'Mixed Use in Theory and Practice: Canadian Experience with Implementing a Planning Principle', *Journal of the American Planning Association*, 68(1), pp. 71–84. doi: 10.1080/01944360208977192

Grant, J.L. (2010) *Good community design: the theory of the public and the practice of the private* (WORKING PAPER: Trends in the Suburbs Project).

GSA (2016) *Demography Survey*.

GSE (2020) 'How Much Area is Required to Set Up a School?' *Global Services in Education,* 2020. Available at: https://www.gsineducation.com/blog/area-required-to-set-up-a-school (Accessed: 15 July 2022).

Gu, Z. *et al.* (2014) 'Circlize implements and enhances circular visualization in R', *Bioinformatics*, 30(19), pp. 2811–2812.

Harland, K. (2008) *JourneyToLearn*. Ph.D. University of Leeds.

Harland, K. *et al.* (2012) 'Creating Realistic Synthetic Populations at Varying Spatial Scales: A Comparative Critique of Population Synthesis Techniques', *Journal of Artificial Societies and Social Simulation*, 15(1), pp. 1–15. doi: 10.18564/jasss.1909

HCDA (2016b) *Population studies of the city of Riyadh 1437AH*. Riyadh.

HCDA (2016a) *The executive plan for coordinating and providing public services in Riyadh: Executive Summary*. Riyadh. Available at: https://www.rcrc.gov.sa/wp-content/uploads/2019/10/005776.pdf.

HCDA (2010) *Coordination plan for the provision of public services in Riyadh: Comprehensive report*. Riyadh.

HCDA (2015) *High Commission for the Development of Arriyadh: Overview* (Accessed: 27 November 2017).

HCDA (2017) *The second operational plan to provide services in Riyadh*. Riyadh.

Heppenstall, A.J. *et al.* (eds.) (2012) *Agent-Based Models of Geographical Systems*. Dordrecht: Springer Netherlands.

Heppenstall, A.J. and Smith, D.M. (2021) 'Spatial Microsimulation', in Fischer, M.M. and Nijkamp, P. (eds.) *Handbook of Regional Science.* Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1767–1784.

Herbert, D. and Thomas, C.J. (1982) *Urban geography: A First Approach*. Chichester: Wiley.

Huff, D.L. (1963) 'A Probabilistic Analysis of Shopping Center Trade Areas', *Land Economics*, 39(1), p. 81. doi: 10.2307/3144521

Huff, D.L. (1964) 'Defining and estimating a trading area', *Journal of marketing*, 28(3), pp. 34–38.

Hunt, J. (2003) *1 DESIGN AND APPLICATION OF THE PECAS LAND USE MODELLING SYSTEM*. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.197.8490.

Hunt, J. and Abraham, J. (2009) *PECAS - For Spatial Economic Modelling: Theoretical Formulation: System Documentation Technical Memorandum 1 WORKING DRAFT*. Template Materials Providing Basic Descriptive Components. Alberta: HBA Specto Incorporated. Available at: https://www.hbaspecto.com/resources/PECASTheoreticalFormulation.pdf (Accessed: 7 September 2022).

Hunt, J., Kriger, D. and Miller, E. (2005) 'Current operational urban land-use–transport modelling frameworks: A review', *Transport Reviews*, 25(3), pp. 329–376. doi: 10.1080/0144164052000336470

Hutchinson, B. and Batty, M. (eds.) (1986) *Advances in urban systems modelling*. Amsterdam: North-Holland (Studies in regional science and urban economics, v.15).

Hyndman, M.R.J. *et al.* (2019) 'Package 'demography''.

Hyndman, R.J. *et al.* (2020) 'Package 'forecast'', *Online] https://cran. r-project. org/web/packages/forecast/forecast. pdf*.

Hyndman, R.J. and Shahid Ullah, M. (2007) 'Robust forecasting of mortality and fertility rates: A functional data approach', *Computational Statistics & Data Analysis*, 51(10), pp. 4942–4956. doi: 10.1016/j.csda.2006.07.028

Iacono, M., Levinson, D. and El-Geneidy, A. (2008) 'Models of Transportation and Land Use Change: A Guide to the Territory', *Journal of Planning Literature*, 22(4), pp. 323–340. doi: 10.1177/0885412207314010

John, P., Dowding, K. and Biggs, S. (1995) 'Residential Mobility in London: A Micro-Level Test of the Behavioural Assumptions of the Tiebout Model', *British Journal of Political Science*, 25(3), pp. 379–397. doi: 10.1017/S0007123400007250

John Bongaarts and Rodolfo A. Bulatao *et al.* (2000) *Beyond Six Billion*. Washington, D.C.: National Academies Press.

Johnson, D.L. (2002) 'Origin of the Neighbourhood Unit', *Planning Perspectives*, 17(3), pp. 227–245. doi: 10.1080/02665430210129306

Jordan, M.I. and Mitchell, T.M. (2015) 'Machine learning: Trends, perspectives, and prospects', *Science (New York, N.Y.)*, 349(6245), pp. 255–260. doi: 10.1126/science.aaa8415

Jordan, R., Birkin, M. and Evans, A. (2012) 'Agent-Based Modelling of Residential Mobility, Housing Choice and Regeneration', in Heppenstall, A.J. *et al.* (eds.) *Agent-Based Models of Geographical Systems.* Dordrecht: Springer Netherlands, pp. 511–524.

Kain, J.F. and Apgar, W.C. (1985) 'The Harvard Urban Development Simulation Model', in Stahl, K. (ed.) *Microeconomic Models of Housing Markets.* (Lecture Notes in Economics and Mathematical Systems). Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 27–71.

Keeble, L. (1952) *Principles and practice of town and country planning*. London: Estates Gazette.

Kennedy, W.G. (2012) 'Modelling Human Behaviour in Agent-Based Models', in Heppenstall, A.J. *et al.* (eds.) *Agent-Based Models of Geographical Systems.* Dordrecht: Springer Netherlands, pp. 167–179.

Keyfitz, N. (1972) 'On Future Population', *Journal of the American Statistical Association*, 67(338), pp. 347–363. doi: 10.2307/2284381

Knaap, G. and Talen, E. (2005) 'New Urbanism and Smart Growth: A Few Words from the Academy', *International Regional Science Review*, 28(2), pp. 107–118. doi: 10.1177/0160017604273621

Knaap, G.-J. *et al.* (eds.) (2007) *Incentives, Regulations and Plans: The Role of States and Nation-states in Smart Growth Planning*: Edward Elgar Publishing (1).

Kornhauser, D., Wilensky, U. and Rand, W. (2009) 'Design Guidelines for Agent Based Model Visualization', *Journal of Artificial Societies and Social Simulation*, 12(2), p. 1.

Krogh, A. (2008) 'What are artificial neural networks?' *Nature Biotechnology*, 26(2), pp. 195–197. doi: 10.1038/nbt1386

Laguna, M. (1997) 'Metaheuristic Optimization with Evolver, Genocop and OptQuest'. Available at: http://aiinfinance.com/metaheur.pdf (Accessed: 25 June 2022).

Laguna, M. (2011) *OptQuest: Optimization of Complex Systems*. Available at: https://www.opttek.com/sites/default/files/pdfs/OptQuest-Optimization%20of%20Complex%20Systems.pdf (Accessed: 27 April 2022).

Laguna, M. and Martí, R. (2002) 'The OptQuest Callable Library', in Voß, S. and Woodruff, D.L. (eds.) *Optimization Software Class Libraries.* (Operations Research/Computer Science Interfaces Series, 18). New York, NY: Springer US, pp. 193–218. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.196.5360&rep=rep1&type=pdf (Accessed: 25 June 2022).

Lapin, M., Hein, M. and Schiele, B. (2018) 'Analysis and Optimization of Loss Functions for Multiclass, Top-k, and Multilabel Classification', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(7), pp. 1533–1554. doi: 10.1109/TPAMI.2017.2751607

Larice, M. and Macdonald, E. (eds.) (2013) *The urban design reader*. London: Routledge (The Routledge urban reader series).

Lee, D.B. (1973) 'Requiem for Large-Scale Models', *Journal of the American Institute of Planners*, 39(3), pp. 163–178. doi: 10.1080/01944367308977851

Lee, R., Carter, L. and Tuljapurkar, S. (1995) 'Disaggregation in population forecasting: Do we need it? And how to do it simply', *Mathematical Population Studies*, 5(3), 217-234. doi: 10.1080/08898489509525403

Lee, R.D. and Carter, L.R. (1992) 'Modeling and Forecasting U. S. Mortality', *Journal of the American Statistical Association*, 87(419), p. 659. doi: 10.2307/2290201

Livera, A.M. de, Hyndman, R.J. and Snyder, R.D. (2011) 'Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing', *Journal of the American Statistical Association*, 106(496), pp. 1513–1527. doi: 10.1198/jasa.2011.tm09771

Lowry, I.S. (1964) *A model of metropolis*. Santa Monica Calif.

Macy, M.W. and Willer, R. (2002) 'From Factors to Actors: Computational Sociology and Agent-Based Modeling', *Annual review of sociology*, 28(1), pp. 143–166. doi: 10.1146/annurev.soc.28.110601.141117

Manson, S. *et al.* (2020) 'Methodological Issues of Spatial Agent-Based Models', *Journal of Artificial Societies and Social Simulation*, 23(1) (23pp). doi: 10.18564/jasss.4174

McFadden, D. (1975) 'The Revealed Preferences of a Government Bureaucracy: Theory', *The Bell Journal of Economics*, 6(2), p. 401. doi: 10.2307/3003236

Middleton, D. (2009) *Growth and Expansion in Post-War Urban Design Strategies: C. A. Doxiadis and the First Strategic Plan for Riyadh Saudi Arabia (1968-1972)*. PhD. Georgia Institute of Technology.

MOE (2000) 'Report of the Kingdom of Saudi Arabia on Childhood End - Decade (1990-2000)'. *The United Nations Special Session on Children*. Available at: https://sites.unicef.org/specialsession/how_country/edr_saudi_arabia_en.PDF (Accessed: 11 April 2021).

MOE (2018) *Indicators of Educaiton*. Available at: https://www.moe.gov.sa/ar/about/Pages/IndicatorsOfEducation.aspx (Accessed: 6 December 2018).

Moeckel, R. *et al.* (2018) 'Trends in integrated land use/transport modeling: An evaluation of the state of the art', *Journal of Transport and Land Use*, 11(1), pp. 463–476. doi: 10.5198/jtlu.2018.1205

Moekel, R. *et al.* (2003) 'Microsimulation of Land Use', *International Journal of Urban Sciences*, 7(1), pp. 14–31. doi: 10.1080/12265934.2003.9693520

MOMRA (2015) *Planning standards for public, regional and local services at their different levels*. Available at: https://www.momrah.gov.sa/ar/node/13020.

MOMRA (2019) 'Requirements for Private Schools'. Available at: http://tbc.sa/Etemad/municipalstipulations.pdf (Accessed: 15 July 2022).

Moreno, C. *et al.* (2021) ‘Introducing the “15-Minute City”: Sustainability, Resilience and Place Identity in Future Post-Pandemic Cities’, *Smart Cities*, 4(1), pp. 93–111. doi: 10.3390/smartcities4010006

Mubarak, F.A. (2004) ‘Urban growth boundary policy and residential suburbanization: Riyadh, Saudi Arabia’, *Habitat International*, 28(4), pp. 567–591. doi: 10.1016/j.habitatint.2003.10.010

Naghavipour, H. *et al.* (2022) ‘Hybrid Metaheuristics for QoS-Aware Service Composition: A Systematic Mapping Study’, *IEEE Access*, 10, pp. 12678–12701. doi: 10.1109/ACCESS.2021.3133505

Nguyen-Luong, D. (2008) ‘An integrated land use-transport model for the Paris region (SIMAURIF): Ten lessons learned after four years of development’, *Documento de trabalho-Institut d'aménagement et d'urbanisme de la région Ile-de-France-IAURIF (não publicado)*. Available at: http://web.mit.edu/11.521/proj08/readings/D_Mes_documentsDNLpredit3ERSA_2008article_SIMAURIF_10_lessons.pdf (Accessed: 29 June 2022).

Ní Bhrolcháin, M. (2011) ‘Tempo and the TFR’, *Demography*, 48(3), pp. 841–861. doi: 10.1007/s13524-011-0033-4

O’Donoghue, C., Morrissey, K. and Lennon, J. (2013) ‘Spatial Microsimulation Modelling: a Review of Applications and Methodological Choices’, *International Journal of Microsimulation*, 7(1), pp. 26–75. doi: 10.34196/ijm.00093

O'Donoghue, C. *et al.* (eds.) (2013) *Spatial Microsimulation for Rural Policy Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg; Imprint: Springer (Advances in Spatial Science, The Regional Science Series). Available at: https://link.springer.com/content/pdf/10.1007/978-3-642-30026-4.pdf (Accessed: 20 April 2022).

Openshaw, S. (1995) ‘Developing Automated and Smart Spatial Pattern Exploration Tools for Geographical Information Systems Applications’, *Journal of the Royal Statistical Society: Series D (The Statistician)*, 44(1), p. 3. doi: 10.2307/2348611

Osman, I.H. and Laporte, G. (1996) ‘Metaheuristics: A bibliography’, *Annals of Operations Research*, 63(5), pp. 511–623. doi: 10.1007/BF02125421

O'Sullivan, D. and Perry, G.L.W. (2013) *Spatial Simulation: Exploring pattern and process*. Chichester West Sussex UK: John Wiley & Sons Inc.

Panella, I., Fragonara, L.Z. and Tsourdos, A. (2021) ‘A Deep Learning Cognitive Architecture: Towards a Unified Theory of Cognition’, in Arai, K., Kapoor, S. and Bhatia, R. (eds.) *Intelligent Systems and Applications.* (Advances in Intelligent Systems and Computing, 1250). Cham: Springer International Publishing, pp. 566–582.

Parker, D.C. *et al.* (2003) ‘Multi-Agent Systems for the Simulation of Land-Use and Land-Cover Change: A Review’, *Annals of the Association of American Geographers*, 93(2), pp. 314–337. doi: 10.1111/1467-8306.9302004

Pearl, J. (1984) *Heuristics: intelligent search strategies for computer problem solving*: Addison-Wesley Longman Publishing Co., Inc.

Perry, C. (2013) ‘The Neighborhood Unit’, in Larice, M. and Macdonald, E. (eds.) *The urban design reader.* (The Routledge urban reader series). London: Routledge, pp. 78–89.

Perry, C.A. (1975) ‘The Neighborhood Unit Formula’, in Branch, M.C. (ed.) *Urban planning theory.* (Community development series, v. 15). Stroudsburg, Pa.: Dowden, pp. 44–58.

Pickardt, T. and Wehrmann, B. (2011) *Land use planning: concept, tools and applications*. Eschborn, Germany: Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ). Available at: https://landportal.org/file/9781/download.

Qhtani, A. and Al Fassam, A.N. (2011) *ArRiyadh Geospatial Urban Information System and Metropolitan Development Strategy for ArRiyadh. Esri International User Conference*. San Diego, July 11-15.

Qurnfulah, E.M. (2015) *The negative impacts of subdivision regulation on the residential built environment: Jeddah's experience*: Newcastle University.

Railsback, S.F. and Grimm, V. (2019) *Agent-based and individual-based modeling: a practical introduction*: Princeton university press.

Railsback, S.F., Lytinen, S.L. and Jackson, S.K. (2006) ‘Agent-based Simulation Platforms: Review and Development Recommendations’, *SIMULATION*, 82(9), pp. 609–623. doi: 10.1177/0037549706073695

Rallu, L. and Toulemon, L. (1994) ‘Period fertility measures: the construction of different indices and their application to France, 1946-89’, *Population. English Selection*, 6, pp. 59–130. Available at: http://www.jstor.org/stable/2949144.

Reilly, W. (1929) 'Methods For The Study of Retail Relationships'. doi: 10.15781/T2XD0RC8G

Rogers, A. (1995) 'Population forecasting: do simple models outperform complex models?' *Mathematical Population Studies*, 5(3), 187-202, 291. doi: 10.1080/08898489509525401

Salvini, P. and Miller, E.J. (2005) 'ILUTE: An Operational Prototype of a Comprehensive Microsimulation Model of Urban Systems', *Networks and Spatial Economics*, 5(2), pp. 217–234. doi: 10.1007/s11067-005-2630-5

Schelling, T.C. (1971) 'Dynamic models of segregation†', *Journal of mathematical sociology*, 1(2), pp. 143–186. doi: 10.1080/0022250X.1971.9989794

Schmidt, B. (2000) *Modelling of Human Behaviour The PECS Reference Mode*: SCS-Europe BVBA Delft. Available at: https://scs-europe.net/services/ess2002/PDF/inv-0.pdf (Accessed: 6 July 2022).

Schoen, R. (2004) 'Timing effects and the interpretation of period fertility', *Demography*, 41(4), pp. 801–819. doi: 10.1353/dem.2004.0036

Schoen, R. (ed.) (2016) *Dynamic Demographic Analysis*. Cham: Springer International Publishing; Imprint: Springer (The Springer Series on Demographic Methods and Population Analysis, 39).

Schwartzman, Y. and Borning, A. (2007) 'The Indicator Browser: A Web-Based Interface for Visualizing UrbanSim Simulation Results', *2007 40th Hawaii international conference on system sciences.* Institute of Electrical and Electronics Engineers, Waikoloa, HI, USA, 03/01/2007 - 06/01/2007. Piscataway, N.J.: IEEE, 92a. doi: 10.1109/HICSS.2007.540

Ševčíková, H. *et al.* (2016) 'Age-specific mortality and fertility rates for probabilistic population projections', in *Dynamic demographic analysis:* Springer, pp. 285–310.

Shang, H.L. and Booth, H. (2020) 'Synergy in fertility forecasting: improving forecast accuracy through model averaging', *Genus*, 76(1) (23pp). doi: 10.1186/s41118-020-00099-y

Shumaker, S.A. and Stokols, D. (1982) 'Residential Mobility as a Social Issue and Research Topic', *Journal of Social Issues*, 38(3), pp. 1–19. doi: 10.1111/j.1540-4560.1982.tb01767.x

Simpson, L. (2012) ‘Demographic change: how planners can prepare for the future’, *The Guardian,* 2012. Available at: https://www.theguardian.com/local-government-network/2012/jul/23/demographics-population-change-planning-future (Accessed: 15 January 2018).

Simpson, L. (2017) *Manual of integrated demographic forecasting for local planning in Wales*. Cardiff: Royal Town Planning Institute. Available at: http://www.rtpi.org.uk/the-rtpi-near-you/rtpi-cymru/policy-in-wales/wales-planning-research-agenda/.

Singh, D., Padgham, L. and Logan, B. (2016) ‘Integrating BDI Agents with Agent-Based Simulation Platforms’, *Autonomous Agents and Multi-Agent Systems*, 30(6), pp. 1050–1071. doi: 10.1007/s10458-016-9332-x

Smith, S.K. and Shahidullah, M. (1995) ‘An evaluation of population projection errors for census tracts’, *Journal of the American Statistical Association*, 90(429), pp. 64–71. doi: 10.1080/01621459.1995.10476489

Smith, S.K., Tayman, J. and Swanson, D.A. (2013) *A Practitioner's Guide to State and Local Population Projections*. (37). Dordrecht: Springer Netherlands.

Sobotka, T. and Lutz, W. (2010) ‘Misleading Policy Messages Derived from the Period TFR: Should We Stop Using It?’ *Comparative Population Studies*, 35(3) (29pp). doi: 10.12765/CPoS-2010-15

Stein, A. (2007) *Modelling qualities in space and time. 5th International Symposium Spatial Data Quality 2007*. ITC, Enschede, The Netherlands, June 13-15. Available at: https://www.isprs.org/proceedings/xxxvi/2-c43/.

Swanson, D.A. and Tayman, J. (2012) *Subnational Population Estimates*. (31). Dordrecht: Springer Netherlands.

Talbi, E.-G. (2009) *Metaheuristics: From design to implementation*. Hoboken N.J.: John Wiley & Sons.

Tayi, G.K. and Ballou, D.P. (1998) ‘Examining data quality’, *Communications of the ACM*, 41(2), pp. 54–57. doi: 10.1145/269012.269021

The state of the world's children 2011: An age of opportunity (2011). New York: UNICEF.

Thunberg, C. and Mannerskog, N. (2019) *Stochastic Gradient Descent in Machine Learning.* KTH Royal Institute of Technology. Available at: http://www.diva-portal.org/smash/get/diva2:1335380/FULLTEXT01.pdf.

Timmermans, H. (2003) *The saga of integrated land use-transport modeling: how many more dreams before we wake up? In Proceedings of the International Association of Traveler Behavior Conference*. Lucerne, Switzerland. Available at: https://archiv.ivt.ethz.ch/news/archive/20030810_IATBR/timmermans.pdf.

Tiwang, R.F. (2020) *A Deep Learning Predictive Model for Source Code Analysis of a Dynamically Typed Language.* Bowie State University.

Tobias, R. and Hofmann, C. (2004) 'Evaluation of free Java-libraries for social-scientific agent based simulation', *Journal of Artificial Societies and Social Simulation*, 7(1).

Torrens, P.M. (2003) 'Cellular Automata and Multi-agent Systems as Planning Support Tools', in Geertman, S. and Stillwell, J. (eds.) *Planning Support Systems in Practice.* (Advances in Spatial Science). Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 205–222.

Torrens, P.M. and McDaniel, A.W. (2013) 'Modeling Geographic Behavior in Riotous Crowds', *Annals of the Association of American Geographers*, 103(1), pp. 20–46. doi: 10.1080/00045608.2012.685047

Train, K. (2009) *Discrete choice methods with simulation*. 2nd edn. Cambridge: New York.

UN (1982) *PROVISIONAL GUIDELINES ON STANDARD INTERNATIONAL AGE CLASSIFICATIONS*. 74th edn. New York. Available at: https://unstats.un.org/unsd/publication/SeriesM/SeriesM_74e.pdf.

UN (2010) *Post enumeration surveys, operational guidelines: technical report*: United Nations.

UN (2019) *World Fertility Data 2019*. Department of Economic and Social Affairs, Population Division (2019). Available at: https://www.un.org/en/development/desa/population/publications/dataset/fertility/wfd2019.asp (Accessed: 4 April 2021).

UN-DESA (2013) *World Population Prospects The 2012 Revision: Volume II: Demographic Profiles*. New York (ST/ESA/SER.A/345).

UNEP, F.A. (1999) *The Future of Our Land: Facing the Challenge Sustainable Management of Land Resources*. Guidelines for Integrated Planning for Sustainable Management of Land Resources. Roma: FAO.

UNESCO-IBE (2011) *World Data on Education: Saudi Arabia*. Available at: http://www.ibe.unesco.org/sites/default/files/Saudi_Arabia.pdf.

UNESCO-ISCED (2012) *International standard classification of education*. Montreal, Quebec. Available at: http://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf.

UNICEF (2000) *The state of the world's children 2000*. (State of the World's Children). New York: UNICEF.

UNICEF (2002) 'The State of the World's Children, 2003'.

UNICEF (2009) *The state of the world's children: Special edition*. New York: United Nations Children's Fund.

Vanella, P., Deschermeier, P. and Wilke, C.B. (2020) 'An Overview of Population Projections—Methodological Concepts, International Data Availability, and Use Cases', *Forecasting*, 2(3), pp. 346–363. doi: 10.3390/forecast2030019

Waddell, P. (2002) 'UrbanSim: Modeling Urban Development for Land Use, Transportation, and Environmental Planning', *Journal of the American Planning Association*, 68(3), pp. 297–314. doi: 10.1080/01944360208976274

Waddell, P. *et al.* (2003) 'Microsimulation of Urban Development and Location Choices: Design and Implementation of UrbanSim', *Networks and Spatial Economics*, 3(1), pp. 43–67. doi: 10.1023/A:1022049000877

Waddell, P. (2011) 'Integrated Land Use and Transportation Planning and Modelling: Addressing Challenges in Research and Practice', *Transport Reviews*, 31(2), pp. 209–229. doi: 10.1080/01441647.2010.525671

Wang, R. (2019) *Lithium-Ion Battery SOC Estimation Using Deep Learning Neural Networks*: Rutgers The State University of New Jersey, School of Graduate Studies.

Wang, R.Y. and Strong, D.M. (1996) 'Beyond Accuracy: What Data Quality Means to Data Consumers', *Journal of Management Information Systems*, 12(4), pp. 5–33. doi: 10.1080/07421222.1996.11518099

Wegener, M. (2004) *Overview of Land Use Transport Models: Handbook of transport geography and spatial systems.*: Emerald Group Publishing Limited.

Wegener, M. (2011) *The IRPUD model*. Dortmund (Working Paper 11/01). Available at: http://www. spiekermann-wegener. com/mod/pdf/AP_1101_IRPUD_Model. pdf.

Wegener, M. (2021) 'Land-Use Transport Interaction Models', in Fischer, M.M. and Nijkamp, P. (eds.) *Handbook of Regional Science.* Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 229–246.

Wegener, M. and Fuerst, F. (2004) 'Land-Use Transport Interaction: State of the Art', *SSRN Electronic Journal (*119pp). doi: 10.2139/ssrn.1434678

Whitley, D. (2001) 'An overview of evolutionary algorithms: practical issues and common pitfalls', *Information and software technology*, 43(14), pp. 817–831. doi: 10.1016/S0950-5849(01)00188-4

Wilkinson, M.D. *et al.* (2016) 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*, 3(1), p. 160018. doi: 10.1038/sdata.2016.18

Wilson, A. (1971) 'A Family of Spatial Interaction Models, and Associated Developments', *Environment and Planning A: Economy and Space*, 3(1), pp. 1–32. doi: 10.1068/a030001

Wilson, A. (2012) *The science of cities and regions: lectures on mathematical model design*: Springer Science & Business Media.

Wilson, A. (2016) 'New roles for urban models: planning for the long term', *Regional Studies, Regional Science*, 3(1), pp. 48–57. doi: 10.1080/21681376.2015.1109474

Winstanley, A., Thorns, D. and Perkins, H. (2002) 'Moving House, Creating Home: Exploring Residential Mobility', *Housing Studies*, 17(6), pp. 813–832. doi: 10.1080/02673030216000

Wiśniowski, A. *et al.* (2015) 'Bayesian Population Forecasting: Extending the Lee-Carter Method', *Demography*, 52(3), pp. 1035–1059. doi: 10.1007/s13524-015-0389-y

Zhang, R., Indulska, M. and Sadiq, S. (2019) 'Discovering Data Quality Problems', *Business & Information Systems Engineering*, 61(5), pp. 575–593. doi: 10.1007/s12599-019-00608-0

Zhao, F. and Chung, S. (2006) *A Study of Alternative Land Use Forecasting Models*. Miami (BD015-10). Available at: https://trid.trb.org/view/784301.