Mobility between Colombian cities is predominantly repeat and return migration

Anonymised manuscript

Abstract

Internal migration is one of the main driving forces of a country's demography. Yet, migration is path-dependent, and critical characteristics of internal migration, including the propensity to keep moving and return to previous locations, are frequently ignored. Here, a model of city-to-city migration is constructed, where the movement of individuals is modelled using the frequency of distinct sequences or signatures. A key novel feature of the model is its ability to account for partial information on an individuals' lifetime migrations. We fit this model to longitudinal data on 3.3 million workers in Colombia, including 1.4 million migrations, and compare signature frequency based on migration and return rates between men and women and between distinct age and income groups. Results show that the majority of people do not move in general, and nearly three out of four times that a person moves at least twice, they return to a previous city. A small group exhibits frequent migration, particularly the young and male. In contrast, women and mature people are less likely to move and more likely to return if they move. At a city level, people from small secondary towns are more likely to leave and not return than people from large metropolitan areas like Bogotá or Medellín.

1 Introduction

People have always migrated. From the first crossing of the Bering Strait to the Spanish conquest, the colonisation of America, the European exodus after wars, migration has always been a central feature of human life (1). Migration is one of the main drivers of the process of urbanisation, industrialisation (2; 3; 4; 5), as well as redistribution of labour resources (6), and of changes in the patterns of human settlement (7). Further, it tends to accelerate ageing in places where the young are more likely to move, increasing the pressure on those who remain and deepening the gender imbalance by altering the gender ratio of a place (8; 9; 10).

We focus on internal migration. The majority of migration is internal (11). Internal migration is a key driver of urban development in many parts of the world (12), and the sorting mechanism of productive workers across cities (6; 13). As countries evolve, increased internal migration is expected (14; 4). In the US, for instance, more than 8 million people move each year from one metropolitan area to another, whilst in the UK, 10% of the population change residence each year (15). Similar patterns are observed in other countries. For example, migration between urban areas accounted for more than half of all internal population flows in Panama, Paraguay and Brazil (16).

Migration is a complex process. Notably, and perhaps unsurprisingly, the longer a person lives in location, the lower the probability that the person will leave their city (17; 18). 'Cumulative inertia', defined by (18; 17) within the context of migration,

refers to both people who move and people who decide to stay in a place. Only a small population group is ever a migrant. Hence, we observe those who frequently keep moving, called "repeat migrants", and those who never move, called "stayers" (19). Often, when a person migrates, they move back to places they lived before. For many reasons, perhaps due to unemployment or to reduce the emotional or monetary costs of being far away from home, people are frequently return migrants (19; 13). Hence, some movements are classified as *onwards migration* while others are classified as *return migration*.

On a collective level, migration is a massive driver of population dynamics. On an individual level, however, migration is a rare event, and hence it is difficult to capture in data. From one year to the next, most people will remain in the same city, and almost no one will be a migrant. In fact, many people will never move, and those who do are likely to remain in their new location for years, or even decades, before moving again. As opposed to alternative types of human mobility, migration occurs over long periods of time, and so even a person who moves often might do it only every few years. Data on migration, however, is rarely collected on such long timescales. For example, administrative data is typically only collected annually or at longer intervals, and rarely spans entire lifetimes. Thus, our challenge is to fully capture an individual and social process that might take decades with only a few years of data containing partial information on both onward and return migrations during a particular time window. We propose a model for migration as a path-dependent complex phenomenon whereby only a few individuals move frequently, and those who do often return to previously-lived places. It enables us to compute the rate at which different population groups move, an aspect of migration that has not yet been fully captured due to the complexity of return and repeated movement patterns and the difficulty of measuring these patterns in real-world data.

To detect repeat and return patterns, instead of observing migration as a one-time type of event, we must think of movements as, for example, migration trajectories (20; 21; 22) which take into account the whole path of a person. Although other migration models have been constructed, it is difficult to detect when a person is returning somewhere if there is an unobserved period during their life trajectory. Here we construct a model with two key components. First, we represent migration trajectories via *signatures* on a network (23; 24). This approach enables us to measure and compare the path-dependent intensity of return and onward migration. Second, we develop a probabilistic model in order to transform the raw data of migration counts into migration *rates* which explicitly take into account the fact that each person is observed only for a finite window of time. Hence, this approach means that we can ignore whether we are observing an individual's first or subsequent migration. We then apply a mixture model to these rates to estimate the number and size of population groups corresponding to distinct rates. This approach fits two model parameters to the observed signatures, the speed (or rate) at which a person migrates (the *migration intensity*) and the speed at which they return (the *return propensity*). The model enables us to reproduce complex migration patterns, and use the adjusted values of the parameters to quantify and compare migration patterns between distinct population groups. For example, we investigate if younger people tend to move more frequently, if women tend to return more and if people with higher income are more likely to stay. Thus, we compare migration patterns by income, gender and age.

Our data source is administrative data on formal employees in Colombia provided by firms to the Ministry of Health and Social Protection. The Planilla Integrada de Liquidación de Aportes (PILA) dataset we use spans 2008-2016 and contains 3.3 million individuals. We use this dataset to detect 1.4 million migrations between metropolitan

areas, smaller cities or the countryside. This data does not capture international migration or internally displaced people due to conflict, which are combined thought to have involved over 10 million people (25). Also, the data only captures formal employees, so there is an unavoidable bias in the dataset.

2 Literature review

2.1 Observing migration through data

Although our understanding of human mobility has drastically evolved over the past few decades due to the availability of geolocated datasets combined with processing power (26; 27; 28), data-based models of migration have been used for decades. For example, the *laws of migration* (29), published by Ernst Georg Ravenstein in 1885, were developed by looking at migration data at a county level from and to the UK, Ireland and Scotland. The law states, among other things, that the majority of migrants move short distances (30; 31; 32), and large towns grow as a result of migration rather than natural population growth (7).

Considering the impact that migration has, data and models of human migration are a valuable tool for forecasting city size and its demand for resources (33; 34; 35). However, migration data is often scarce and challenging to manipulate (36). Census data is often used as a source of migration data, typically including information on the place of residence five years before the census, one year prior, and at the time of the census (37; 38). However, census data has many limitations, including the time period between waves and that it can only capture a limited number of migrations. Thus, census data is frequently combined with other sources (39; 40; 41; 42; 43). Besides census data, population registers and administrative datasets have become a valuable source of migration data (44). However, these data types are not available in most countries, so other data sources are often exploited.

Various novel data sources that capture more dimensions of migration (beyond census data) have been used to analyse human mobility and migration. Migration is a specific type of human mobility, often defined as a change of residence for a period of at least six months or one year (45; 46). Thus, there is some overlap in the types of data and techniques to model migration and daily mobility. These include, for example, data from social media (36; 47; 48), card transactions (49), mobile phone data (50; 51; 52; 53), and others (54; 55; 56). Using a combination of mobile phone call records and manual data collection, researchers have analysed the time spent in a location and the probability of returning (53), classified individuals as explorers and returners (54), and showed that daily trips can be described by a limited number of patterns (57). It is also possible to use social media to obtain up-to-date demographic estimates, and nowcast migrant stocks (36; 47).

One of the most significant challenges for the study of migration is that mobility data is often a byproduct of sources that were not originally intended to provide information on the movements of people (58). Migration data is challenging, with some significant details often unknown. For instance, individuals are commonly observed from an arbitrary starting point (particularly when individuals are not surveyed), which means that their precise origin (and likely first migrations) are ignored. Thus, it is impossible to detect if a person is returning as the data might not capture previous locations. Many migration studies are based on surveys, with different waves of individuals each time, and

so migration is often treated as a one-time event (59; 21). Hence, detecting if a person moves frequently is challenging.

2.2 Repeat, return and onwards migration

Longer residence in a location reduces the probability that the person will move (17; 18; 60), whilst recent migrants often migrate again (61; 62). A person who recently moved has, in general, more information about the moving process (59) and has fewer ties with the new location (38) and so they might have fewer reasons to stay. Inertia is observed both for people who move and for people who decide to stay in a place, meaning that the population of a country can be divided into those who frequently keep moving and those who never move (19). Repeat migration is a pattern that has been observed for decades (63; 64). Focusing on a small US town, for example, a study found that the departure rates of new arrivals was more than twice that of long-term residents in the 1930s and 40s (65). Most migration is done by a small group of "hypermobile" people who change residence repeatedly and frequently (66).

Not only are past migrants more likely to keep moving, but they will also likely return to previous locations (67). We have a high propensity to return to places we have visited before (53). Movements of people are classified as return migration (if they go back to any location in which they previously lived) and onwards migration (if the location is new for the person). Return migration might be motivated by costs (emotional or monetary), perhaps as expectations of moving were not fully satisfied or due to predetermined intentions of going back (39; 38; 19). Return migration allows individuals to use information about previously known locations, reducing uncertainty. In some African countries, for example, half of the men moving from urban to rural areas are return migrants who lived in rural areas as children (40). At a country level, return migration is increasing in Mexico, China, and others (45), with significant consequences. In Mexico, for example, households with return migrants exhibit a significantly higher school attendance (68).

The propensity to return decays with the length of absence (67). Repeat and return migration are frequently observed patterns that are difficult to capture in data (61; 39; 4). Particularly when individuals are not surveyed, people are observed from an arbitrary starting point in their lives, so their first locations and ties to the place where an individual grew up may not be captured (38). In previous related work using daily mobility data, individuals were grouped into distinct profiles (explorers and returners) characterising their mobility patterns (54), where returners limit much of their mobility to a few locations. Here, in our work, individuals are classified according to their migration patterns into groups depending on how frequently they move and how often they return to previously-lived locations.

2.3 Characteristics of those who move and those who stay

Besides migration being rare, complex, and path-dependent, certain individual characteristics greatly influence the chances of staying or moving and returning. Mobility rates decline with age (69; 16; 70), and vary according to life events such as marriage, family formation and retirement (71; 72; 73). Also, the age of a person at their first migration tends to have a high impact on their life-course trajectory (74; 75). Besides age, gender also matters (76; 77). For example, the reunification of Germany attracted many young women from East and West that led to a tremendous deficit of women in Eastern Ger-

many (8). There are some differences in terms of mobility between men and women. For example, women tend to make shorter trips (78) and visit less diverse places than men (79). However, differences between men and women tend to be smaller in richer countries (40).

In addition to age and gender, other factors influence migration propensity such as income (76) or risk aversion (80). For example, single and childless individuals migrate more often (22). Also, those who often moved as children are more likely to move as adults (81; 82), so migration can be thought of as a learned behaviour (42). Mobility is a cumulative process that takes place over the entire life of a person (82).

2.4 Modelling migration

Understanding migration is difficult since there are many individual and collective factors that affect it, including effects of physical distance (83; 84) and a non-linear impact of origin city size (85). There is no perfect data that captures migration, but also, migration is challenging from the modelling point of view. Among the many techniques, networks are a natural way to model human mobility since distinct locations can be summarised as nodes, and movements can be represented by connections between nodes. Network analysis is a powerful tool to analyse spatial and temporal data, as is the case with migration (86). People move on a spatial structure which can be considered as a spatial network (87).

Mobility networks have been used to study migration (88; 89; 19), life trajectories (22; 21) and daily human mobility (87; 57). In terms of migration, distinct locations (cities or municipalities, for example) could be the network nodes, and the weights of the edges represent the frequency of journeys or the rate of migration between two locations. This type of network has been used to model migration between California and other states in the US (89; 88) or people moving from, and to Germany (19). Also, when the total outflow and inflow of each node are known, the entropy maximising spatial interaction model produces estimates of the flows between geographical locations (90). Based on a network, it was found that most mobility patterns can be described by a very reduced number of daily networks or "signatures" (57). Network signatures are subgraphs that occur more often than would be expected in random networks (24), and so they are one method to describe human mobility (57).

When an agent moves in a network, they define a sequence of visited nodes. Therefore, instead of considering the underlying network, often models are based on the sequence of nodes only. Sequence-based methods treat trajectories as the unit of analysis, facilitating the identification of patterns in temporal sequences. For example, cities can be classified depending on their size, say A for the smallest cities and I for the largest ones. The sequence BGI is the trajectory defined for a person who moved from a small (B) to a medium (G) and then to a large city (I). Counting repetitions of sequences is a powerful technique to capture how frequently people move between cities of different sizes (20). Similar methods have been used to model career pathways (91). Also, comparing trajectories of the zones visited in a city, it was observed that women made more self-loops, suggesting different mobility patterns by gender (78). Sequence analysis helps us summarise multiple migrations, including repeats and returns, as a simple string of characters (59; 22).

3 Methods

3.1 Defining metropolitan areas and countryside

With more than 80% of its 50 million inhabitants living in one of 62 geographically isolated cities, Colombia is representative of highly urbanised middle-income Latin American nations and presents an ideal observatory from which to observe and analyse inter-city migration patterns. Bogotá is the capital and largest city of Colombia, with nearly 10 million inhabitants in its metropolitan area, followed by Medellín (3.9 million), Cali (3 million) and Barranquilla (2.4 million inhabitants). Its territory encompasses parts of the Amazon rainforest, the Andean highlands and deserts, so the country has rich and diverse geography at the cost of long intercity distances. For example, the Euclidean distance between the two major cities, Medellín and Bogotá is 240 kilometres, but it takes more than eight hours to drive between both cities on a winding road that is 74% longer than the Euclidean distance. This paper uses data from the Colombian Social Security (PILA), which contains nine consecutive years of administrative data for 3.3 million formal employees between 2008 and 2016. Within this time period we extract 1.4 million city-to-city migrations.

Our data provide the municipality where the employee is working. Colombia is divided into 1,122 municipalities, some with a population of a few million and the three smallest municipalities with less than 1,000 inhabitants in 2018. Some of the municipalities are part of the same metropolitan area, such as Medellín and Envigado, and many of the municipalities are rural. We merge urban municipalities into metropolitan areas, spatially delineated through commuting patterns, according to previous research (92; 93).

The metropolitan area of Bogotá, for instance, consists of 23 municipalities, including the municipality of Bogotá itself, Soacha, Factativá and others. In total, 19 metropolitan areas are formed as the union of two or more municipalities (the four largest metropolitan areas are Bogotá with 9.7 million inhabitants; Medellín with 3.9 million; Cali with 2.9 million; and Barranquilla with 2.4 million inhabitants) and are labelled with the name of the largest municipality. Movements of people within the same metropolitan areas are ignored (for example, if a person moved from Soacha to Bogotá it is not considered migration).

Also, 43 municipalities that are not part of the 19 metropolitan areas but are urban are considered separate "cities", including Ibagué and Santa Marta (with more than 500 thousand inhabitants). In total, we obtained 62 "cities" with this method (19 metropolitan areas and 43 urban municipalities) and used the term cities to refer to this set. We only have the municipality of the person, which is challenging for less urbanised areas. Some municipalities are very large. For example, Cumaribo in Vichada has more than 65,000 square kilometres, larger than Sri Lanka or Costa Rica, and movements inside municipalities are impossible to trace. Therefore, municipalities that are not added by this method are considered to be "countryside".

In total, 964 municipalities are labelled as the countryside and movements between different parts are not considered, although movements between the countryside and cities are studied.

Migration between 63 distinct locations is studied: 19 metropolitan areas; 43 urban municipalities, and the countryside, corresponding to the address of the person's job.

Observation	Kept?	Imputation	Signature
$\circ\circ A\|AAA\|AAA$	No	–	–
$AAA\|\circ\circ B\|BBB$	No	–	–
$AAA\|AAA\|A\circ\circ$	No	–	–
$\circ AA\|A\circ A\|A\circ A$	Yes	$AAA\|AAA\|AAA$	A
$AAA\|\circ BB\|BBB$	Yes	$AAA\|ABB\|BBB$	AB
$\circ A\circ\|B\circ C\|\circ D\circ$	Yes	$AAA\|BBC\|CDD$	$ABCD$
$AB\circ\|CD\circ\|A\circ E$	Yes	$ABB\|CDD\|AAE$	$ABCDAE$

Table 1: Distinct scenarios on the filtering and the imputation based on the observed data for each year. The letters $A, B, C, \ldots$ represent different cities, and $\circ$ represents a year with unknown location.

3.2 Data selection and missing information

There are 16,576,254 observations in the PILA dataset. Any formal employee who worked for at least one month in any given year between 2008 and 2016 appears on the dataset. For each year, the most frequent location of each person is selected. For some individuals, there is some missing data, which includes, for example, people who retired between 2008 and 2016, a person who did not work for a period of time or who joined the labour market after 2008. It is impossible to detect migration and compare individuals whose location is known only for a few years. Therefore, we have two strategies to obtain as many comparable observations as possible: we filter out individuals for whom not enough information is known (so some of their migrations might be missed), and we impute the missing data for the kept individuals.

The procedure is as follows:

- Observations are dropped if they have any two consecutive years of missing information.
- For the remaining individuals, if there is one missing year, the missing location for that specific year is imputed by the location of the previous year.
- If an individual is missing the location of the first year, it is imputed by the location of the second year.

Schematically, if we represent with $\circ$ the missing year, we have the following filtering and imputation scenarios (Table 1).

With this filtering and imputing process, individuals who could have lived in a city for more than one year without it being detected in our dataset are dropped (as the first three examples on the table, which could be an undetected migration), but we keep individuals for whom undetected migration is not possible. The imputing procedure does not increase the number of migrations and does not alter signatures. For instance, on the fourth row of Table 1, there are three missing years, and they are all imputed with A, which assumes that the person did not move, or at least, not for a sufficient time to be considered a migration. In the fifth row, for which there is one missing year, this alternatively could be imputed with A or B, but it would not change the number of migrations or locations of the individual or the migration signature. An individual is kept if they are missing up to five years, but those are not consecutive (as the sixth row), and the procedure also keeps return migrations (as the seventh row of the table).

International migration cannot be traced using administrative records. Suppose a person does not appear in the dataset for a couple of years, for example. In that case, it could be due to unemployment, informal employment or international migration, and so they are treated equally. People from other countries without a formal job are not identified in the dataset either. Only if they belong to the formal sector are traced, and their movements are quantified identically as every other person.

3.3 Representing onward and return migration

The frequency of specific patterns of onward and return migration is captured by constructing the sequence for each individual. Schematically (as in (59; 20)), we represent the *known locations* of an individual across the nine years we observe them in our dataset as a sequence of nine characters. For example, we could have $AAA|AAA|AAA$, $AAA|AAB|BBB$, or $AAA|ABB|CCA$, where A, B and C represent different any of the 19 metropolitan areas, 43 cities or the countryside in Colombia, which is also represented by a single letter, giving us 63 distinct locations. From the sequence of 9 characters, we remove repetitions of consecutive locations and obtain the *signature*. This is a sequence where all individuals start at A, and if they move to new locations, they move in alphabetical order (so they move to B, then C and so on). If an individual does not move, we get simply A. If an individual moves once, we obtain signature AB. If a person, for example, moves for a second time, then return migration forms the signature ABA, and onward migration forms the signature ABC. Some further examples of signatures are shown in Table 2.

Known locations	Signature	Migrations M_i	Locations L_i	Number of returns	Number of onward
$AAA\|ABB\|BBB$	AB	1	2	0	1
$AAB\|BBB\|BAA$	ABA	2	2	1	1
$AAA\|ABB\|CAA$	$ABCA$	3	3	1	2
$AAB\|ACC\|CCB$	$ABACB$	4	3	2	2
$ABC\|CDD\|BAC$	$ABCDBAC$	6	4	3	3

Table 2: Known locations, signatures, number of migrations, distinct locations, return migrations and onward migrations based on the person's imputed data.

The idea is to analyse the observed frequency of different signatures in the data. The process of reducing the known locations of a person into a signature and then analysing their frequency has been used in mobility studies (94), where the set of locations forms a network. Thanks to signature analysis, it was detected that 90% of daily mobility can be described by 17 distinct signatures (57). Here, we observe that many individuals do not move during the nine years of the data, but some move more than once according to different signatures. In particular, signatures which include a return to a previous city are very frequent. For example, from the group of individuals who moved twice (8% of people), it is observed that 81% of them returned to their previous location and only 19% exhibited onwards migration. And individuals who moved three times are 2.9 times more likely to have moved only between two cities than between four cities. Overall, we find that 95% of individuals who move can be described using signatures that include just three distinct characters, such as AB, $ABAC$, $ABCAB$, $ABABC$. Thus, there is a very high rate of return migration (Figure 1). Other, more complicated signatures are

Figure here

Figure 1: Distribution of signatures when the person moves four times or less. All individuals begin at the blue node (left) and move according to the arrows. When a person moves twice, for instance, the resulting signatures is either ABA, which represents 6.51% of the observed signatures, or ABC, which represents only 1.52% of the signatures.

also observed. For example, $ABCADA$ was observed among 0.19% of individuals. In total, we encountered 20,374 different signatures in the dataset. See the Supplementary Information for the frequency of the top 30 signatures.

Below we propose a parameter-based model to generate a similar frequency of distinct signatures as the observed ones, similar to a model designed for capturing daily mobility patterns (53). We then fit these parameters to the data in order to empirically capture the return rates and the concentration of migration for particular sub-populations.

3.4 A migration model based on signatures

Migration is a rare event. From one year to the next one, roughly 94.5% of the population remains in the same city. If migration after one year is independent of previous years, and if we randomly choose 94.5% of the population to remain in the same city and let 5.5% move, then after eight years, 63.7% should still be in the same location - but data shows that 77.2% of people are still in the same city. Too many people are more likely to remain than what we would observe if individuals are randomly picked each year. Therefore, migration is indeed path-dependent as moving is affected by previous decisions and moves. In turn, some people move frequently, and others remain, so we observe an inhomogeneous migration pattern. We model this inhomogeneous pattern by dividing the population into sub-population groups (see the Supplementary Information for more details).

Internal migration is a rare and highly concentrated social event, and so here we apply a technique that has been applied to crime data, where victimisation is also rare and concentrated (95; 96). In terms of crime, people are usually observed via yearly victimisation surveys. If a person was not the victim of any crimes for a period of a year, it does not imply that the person is immune to suffering crimes (95). Events that happen with a small frequency at an individual level, such as suffering a crime or moving between cities, should be analysed as a rare event, that is, based on the rate or speed at which they happen. That rate might be small, for example, a person might move between cities every few years, and that low rate is precisely what we want to capture.

One of the problems with data collection in a passive manner is that we begin observing a person at some arbitrary starting point. This means that we do not know if their first known location corresponds to their origin, or even if their first (known) migration is a return migration to a city in which they previously lived or an onward migration. Instead of assuming that people have not moved before we begin observing them, we propose a method to estimate the *migration rate* of each person and then group individuals based on their rate. Formally, let $M_i(t)$ be the number of times that person i has moved between time t_0 and t. Let us assume that if a person moves, it does not affect the probability of future migrations, but rather, moving frequently is the result of a high migration rate. Hence, migrations occur independently, and we assume that the person's rate is constant. Thus, if someone moves more frequently than another, it is due to a

difference in their migration rates. Therefore, the number of migrations follows a Poisson distribution with rate $\lambda_i t$,

$$M_i(t) \sim Po(\lambda_i t). \tag{1}$$

Both are strong assumptions with respect to migration (independence of observations and constant rate) that are not observed in reality (71). It is known that the longer a person lives in a single location, the lower the probability that the person will leave their city (17; 18). Hence, a migration is more likely to occur after a recent migration (not independence), and a constant rate is not necessarily the case (66; 62; 42). However, for short periods of time, regarding migration as a Poisson distribution, assuming a constant rate, enables us to analyse the rate at which individuals move (λ_i) rather than the number of migrations directly. Assuming a constant migration rate is problematic for extended periods.

The parameter $\lambda_i \geq 0$ is a rate, or speed, at which the person i moves. We can think of the function $f(t) = E[M_i(t)] = \lambda_i t$ as a straight line which indicates, for any $t \geq 0$, the expected number of migrations of person i. Only if $\lambda_i = 0$ then the function $f(t)$ is a horizontal line, indicating that the person will not move. Otherwise, even if the gradient of f is small, the person i expects to eventually move. This approach enables us to take into account the fact that migration is a rare event, meaning that even if a person did not move during the period we observe them, they might eventually move. Hence, instead of counting the number of migrations directly, we observe migration through the lens of "speed", and that speed can be small. In turn, issues with the arbitrary starting point and observation window are less relevant.

Let $L_i(t)$ be the number of distinct locations that individual i has lived in since time t_0. Then, if $M_i = 0$ (the person has not moved) the number of locations is $L_i = 1$. With $M_i = 1$, then $L_i = 2$, since the person moved between two locations. But, if $M_i = 2$, then on the second migration the person might have moved to their first location (ABA), or might have moved to a new location (ABC). If a person has lived in many cities, then it is more likely that they return to one of these cities (relative to someone who has not frequently moved before). Since the number of cities is large relative to the number of moves, instead of integrating this increased probability into the model, we simplify it and assume a fixed probability of returning. Thus, assume that for each migration after the first one, the person decides whether to move back to a previously known location, with probability π, called the "return rate", or moves to a new location with probability $1 - \pi$. Then, the conditional distribution of $L_i(t)$ given $M_i(t) = m$ is given by

$$L_i(t)|[M_i(t) = m] \sim Bin(m - 1, 1 - \pi) + 1, \tag{2}$$

if $m > 1$, so the person moves more than once, and

$$L_i(t)|[M_i(t) = m] = m + 1 \tag{3}$$

if $m = 0$ or 1, so the person does not move, or moves only once.

It is easy to show that a Binomial distribution, conditional on a Poisson distribution, also follows a Poisson distribution, with the combined rates λ_i and probability π, so that

$$L_i(t) \sim Po(\lambda_i(1 - \pi)t) + 1, \tag{4}$$

if the person moves at least once.

The novel aspect of our method is that instead of migration rates based directly on the observed data, we look at the *speed* (or rate) at which migration occurs λ_i. Since

Figure here

Figure 2: The model has two parameters, λ (horizontal axis), which represents the rate at which a person moves, and the return rate π (vertical axis), which is the propensity that a person returns to a previous city. Different signatures are obtained for distinct regions of the parameter space. A high migration rate and high return rate, for example, corresponds to signatures such as $ABABCAB\ldots$ (which are long sequences with a small number of distinct characters), a medium migration and a medium return rates represents shorter signatures with fewer repetitions such as $ABCDC$, while a low migration rate corresponds to, e.g., A or AB.

migration is a rare event, the estimation is based on the fact that it might take years for a person to move. Further, a person might return (with even smaller frequency) to previous locations, so we discount the migration rate λ_i by the return rate $1-\pi$ and obtain the speed or rate at which a person moves to new locations, $\lambda_i(1-\pi)$. Both rates enable us to describe and compare migration patterns in a succinct manner (Figure 2). The method enables us to extend the observed patterns outside the limits of the data by computing the expected number of migrations and locations for each person.

3.5 Reproducing complex signatures

Internal migration patterns are complex, and they are the emergent result of millions of people deciding to move for personal reasons but exhibiting collective behaviour. A generative algorithm is constructed for the sequence of cities of each individual, which takes as input the migration signature of each person and a model parameter called the *return rate*, which is estimated as follows. Firstly, individuals are grouped based on a mixture model into an unknown number of groups, each group of individuals with the same migration rate λ_j. A sampled probability $\pi \in [0,1]$ is used to simulate signatures. For person i, with migration rate λ_i, the number of migrations in τ years is sampled from $M_i \sim Po(\tau\lambda_i)$. After the first migration, and each time a person moves, they move back to any of the cities in which they previously lived with probability π. If there is more than one city in which they previously lived, it is randomly chosen. The algorithm generates a sequence of M characters, which is then transformed into a (simulated) signature. For a population of P individuals, we compare the observed frequency of observed and simulated signatures and keep the value of π which produces the best fit to the data ($\hat{\pi}$, i.e., the value which best reproduces the observed frequency of signatures).

We also compare the frequency of each signature with two distinct models. A null model, with no return migration, so $\pi_H = 0$, and so each time the person moves, they choose a different city with no return migration, so signatures arise such as A, AB, ABC and so on. We also compare with a second model with complete randomness in the sequence of cities, so each time the person moves, they select any of the $N-1$ cities randomly and moves. Under complete randomness, return migration could occur after the first migration if the person randomly chooses a city in which they previously lived. If $\hat{\pi}$ gives values close to zero, we observe discouraged returns, meaning that individuals rarely move back. In the extreme case, if $\hat{\pi} = 0$, we obtain the model with no return. With values of $\hat{\pi} \approx 1/(N-1)$ we observe random returns, meaning that return migration happens at a similar frequency in which randomness would happen. With high values of $\hat{\pi}$ we observe preferential returns, meaning that people return more frequently than

Figure here

Figure 3: The horizontal axis is the observed frequency of signatures (in a logarithmic scale), and the vertical axis is the modelled frequency. The yellow line is the identity, and observations closer to the line have a better fit. The size of each mark is proportional to the frequency so that more frequent signatures have a bigger mark.

randomness to previous cities, where the extreme case of $\hat{\pi} = 1$ produces only signatures with two cities, with the form $ABAB\ldots$.

After estimating the best fit of $\hat{\pi}$, we observe that the model with a return rate $\pi_H = 0$ is capable of dividing those who move and those who do not move and produces some of the onward migration patterns (AB, ABC and so on). However, it tends to overestimate their frequency (Figure 3). The return model (with a return rate $\pi_R = 1/(N-1)$) captures some of the frequency of return migration, but it overestimates the signature ABC and underestimates the signature ABA, as it does not favour return migration. This behaviour is best captured by our model, named the Poisson return model, which mimics the signatures generated by migrants in Colombia quite well. For the frequently observed signatures, the Poisson return model does capture most of the frequency, so the model can reproduce the frequency of complex signatures, such as $ABCADA$ or others.

The most frequently observed signatures are captured with the Poisson return model with a value of $\hat{\pi} = 0.7612$, with a return rate much higher than the random model $\pi_R = 0.0161$, meaning that return migration happens much more frequently than under a model assuming random movements. Return migration is indeed much more frequent than a null hypothesis, and people are much more likely to go back than move to new locations. The Poisson return model is also a generative algorithm since it is possible to simulate migrations for longer (or shorter) periods of time.

4 Results

4.1 Classifying individuals based on their migration rate

The individual migration rate λ_i and the return rate π enables us to quantify and reproduce migration signatures using a minimal number of parameters. This, in turn, enables us to compare migration rates between distinct populations, such as men and women, or between the young and mature. We combine individuals using a mixture model (97), a novel technique in mobility studies, which groups individuals based on their rates. This technique is frequently used in medical studies to divide populations into distinct groups (97; 98), and has also been applied to study victimisation rates. In this case, instead of migrations, the number of crimes suffered by each person, and its associated rate, is analysed (96; 95).

Using a mixture model, we group individuals based on their migration rate into k groups, such that $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_k$ and with relative sizes $q_1, q_2, \ldots, q_k$ such that $\sum q_j = 1$. All individuals are assigned to a single group with the same migration rate. The number of groups k is determined by the data, and not known *a priori*. In this case, distinct groups may correspond to 'types' of individuals in terms of migration, i.e., those who move frequently (or rarely) will be grouped together.

For a population, its corresponding parameters are estimated as follows. We fit a finite mixture model (99; 97; 100) which takes as input the number of migrations of each

individual, $M_1, M_2, \ldots$, and gives the number of groups, $\hat{k}$ and the corresponding rate $\hat{\lambda_1}, \hat{\lambda_2}, \ldots, \hat{\lambda_k}$ and size $\hat{q_1}, \hat{q_2}, \ldots, \hat{q_k}$ of each group. Then, the best fit return rate π is estimated from the data by generating a simulated sequence of nine characters (which represents a signature for a person), and minimising the mean square error between the observed and simulated frequencies for a range of π. The return rate is also compared to scenarios with discouraged ($\pi = 0$), preferential ($\pi = 1$) and random returns ($\pi = 1/N$), where N is the number of locations. Results show that the models with no return migration and with random returns cannot reproduce the observed frequency of signatures (see the Supplementary Information for more details).

Hence, with just a small set of parameters (the migration rate of the group λ_j, its size q_j and the return rate π, which is assumed to be the same for the whole population), we can summarise movements and measure return migration for distinct population groups. Once the parameters are obtained, after a period of τ years, a person expects to move $\tau\lambda_P$ times and expects to live in $\tau\lambda_P(1-\pi_P)+1$ distinct cities for values of $\tau > 0$.

The idea behind the model is that we assume that some people will frequently migrate whilst some remain. The Poisson distribution captures the general trend, whereas observed variations in the number of locations around the expected value are unobserved context-specific aspects of each person.

4.2 Most individuals will never migrate

The mixture model applied individuals in our dataset yields $\hat{k} = 3$, meaning that based on the number of migrations, people can be divided into three groups (relative sizes and rates are shown in Table 3). Our results show that 69.3% of people will not move and can be considered as "stayers", and that there is a small group containing 2.5% of the individuals who move very frequently, at a rate of $\hat{\lambda}_3 = 3.04$, who can be considered "supermigrants".

	Group	$\hat{q}$	$\hat{\lambda}$	$\hat{\lambda}(1-\pi^\star)$
1	stayers	0.6926	0.0000	0.0000
2	migrants	0.2826	1.2827	0.3077
3	supermigrants	0.0248	3.0331	0.7275

Table 3: Migration profile of Colombia. The finite mixture model divides the population into three groups: one containing nearly 70% of the population that does not move, and one with less than 2.5% of the population that moves very frequently.

Data shows that 77.22% of the population does not move from their location between 2008 and 2016. Yet, when we divide the population according to their migration rate, only 69.26% are considered stayers (so their rate is $\lambda_1 = 0$). The difference comes precisely from observing migration as a rare event. A person from group 2 does not move with a probability of $\exp(-1.2827) = 0.277$, meaning that even when people from that group are expected to move, 27.7% of that group will not move inside the window in which we observe them. Similarly, for group 3, where 4.8% are expected not to move, but they eventually will move (outside the window we observe them) as opposed to the stayers. It is by observing the gradient at which people move instead of the counts directly that we can infer data outside the nine years that our observation window lasts.

The rate of return migration is $\hat{\pi}^\star = 0.7612 \pm 0.0007$ (with intervals obtained through bootstrapping), meaning that after the first migration, roughly three out of four times a

Figure here

Figure 4: Square error of the signature frequency (observed vs simulated) for different values of the return rate π, where the value $\pi^\star = 0.7612$ minimises the error. The same analysis (and corresponding curves) are computed for the return rate for male and female, and similar (curves nor shown) for young and mature populations. Random returns would be when $\pi = 1/62 = 0.0161$, so there is a much higher preferential return, particularly for women.

person moves back to a previous city. Our results indicate that there is strong preferential return migration, far from the no-return values with $\pi_H = 0$, and from random returns $\pi_R = 0.016$ but also far from the case with only returns $\pi_O = 1$ (see the Supplementary Information for the details). The distribution of the number of distinct locations of a person in group 3, the supermigrants, for instance, is $L_3 \sim Po(0.7275) + 1$, which means that roughly 18% of the population from that group expects to live in three or more cities.

As with wealth or other individual-level variables, migration is highly concentrated. A few individuals earn more money than thousands or even millions of people combined, and similarly, a few individuals have a higher rate of migration than the vast majority of the country. We can deploy this methodology to uncover migration rates for different population partitions, e.g., men and of women; low and high income and young and elderly, along with their return rates π_m and π_w respectively. This enables us to shed light on distinct migration behaviours observed for sub-population groups.

4.3 Return rate for different subgorups

The rate of return migration for the whole population is $\pi^\star = 0.7612$, obtained by minimising the mean square error between the observed and the modelled signatures using the Poisson return model (Figure 4). For other subgroups, the return rate, estimated with the same generative algorithm, has different values. The smallest return rate is obtained for the 25% youngest population with $\pi_Y = 0.5799$ and the highest is obtained for women $\pi_W = 0.8161$, so that women are 40% more likely to return than a person from the 25% youngest group.

4.4 Women move less and return more frequently

We divide the population by gender and consider each group separately. In the data, 61.4% of individuals are male, and 38.6% are female. For both genders, the mixture model gives $\hat{k} = 3$ groups, meaning that both men and women can be subdivided into three groups. The *migration profile* of each gender (or any subpopulation) is determined by the set of pairs (q_j, λ_j) for each group. These pairs provide an overall description of the migration intensity and can be visually displayed as a step-wise function (see Figure 5 in which individuals are plotted on the horizontal axis, and their corresponding rate on the vertical axis). We observe that both men and women have a sizeable group who is a stayer (65.8% in the case of males and 74.5% in the case of females), and in the case of men, a small group (4.2%) are identified as supermigrants. The average migration rate for men is $\lambda^{(m)} = 0.53$, and the average migration rate for women is $\lambda^{(w)} = 0.29$, meaning that, as it has been observed elsewhere (101; 102) and for other types of mobility (79), men move more frequently than women.

Figure here

Figure 5: Migration profile for women (upper part) and for men (lower part) in the upper panel, where all individuals are sorted (horizontal axis) according to their migration rate (vertical axis). A similar comparison in the middle panel for young and mature populations. In the bottom panel, the migration profile of the population with lower and higher income.

Figure here

Figure 6: Observed migration signatures for women (top panel) and for men (bottom panel). Women who move twice, for instance, return to their first city (ABA) 84% of the time, and move to a new city (ABC) 16% of the time. Men return to their first city 79% of the time and move to a new city 21% of the time.

Unlike what was encountered for international migration, where it was observed that men are more likely to move and return than women (77), we find that women are more likely to return ($\pi^{(w)} = 0.8161$) than men ($\pi^{(m)} = 0.7407$). Men are more likely, however, to explore a new city. Comparing, for instance, the most migratory groups between men and women, the expected number of distinct locations is 45% larger for men than for women. The signatures per gender (Figure 6) highlight that women are more likely to stay in the same city, to move only once, if they move, to travel slightly shorter geodesic distances (5% shorter on average). If they move twice or more, they are more likely than men to return to previous cities.

The migration profile is a multi-dimensional description of migration rates at a population level. However, given two distinct migration profiles (say, men and women), it might not be easy to detect if the rates are more or less concentrated across the population. The *concentration of migration* is computed as the Gini index of the migration rates from the population (96). Since the Gini index has scale independence (some population groups could be more migratory as a whole) and population independence (it does not matter how large the population is), it is a comparable metric for the concentration of migration. We find that the concentration is higher for women (0.75) than for men (0.69). This is because women move at a lower rate than men, and those who move do it at a lower rate with a higher chance of returning than men.

4.5 Mature and low-income people move less and return more

Age is a key driver of migration (77), whereby young adults have the highest migration intensity (101) but with significant variation according to age (103). For example, younger adults tend to move to large cities to benefit from density, jobs and amenities, but elderly people frequently move in the opposite direction (70; 72; 73; 84; 101). To investigate these differences in our dataset, as we did with gender, we subdivided the population into two according to their age in 2016. We designate the 25% youngest population (age smaller than 35.9 years) and the 25% oldest population (age larger than 50.9 years) as the *young* and the *mature* population respectively (Figure 5). The young population is subdivided by the mixture model into three groups, where 62.9% of the population is a stayer and 1.8% is a supermigrant, with a rate of $\lambda_3^{(y)} = 3.2$. The mature population is subdivided into four groups, where 76.1% of the population is stayer, and 2.2% of the population is supermigrant, with a migration rate $\lambda_4^{(r)} = 3.1$.

Although the young and the mature population have nearly the same size supermigrant group, with a similar rate, the young population is nearly twice as migratory as the mature population and has a more explorer profile as has been observed for international migration (77). Combined, the young population expects to live in 16% more distinct locations than the mature population. Unlike what we encountered with respect to the gender divide, the distance between origin and destination for the young and the mature populations is of similar magnitude (223 kilometres each time a person moves).

It is frequently assumed that income or employment are some of the main drivers of migration (104), whereby a person compares their expected income with and without moving (and among distinct destinations) and chooses the best option (105; 106; 30; 27). Migrants are frequently thought to be 'pushed' out of areas with lower income and attracted to areas with higher earnings (107; 108). Still, as it has been noted, expectations might not materialise, or people could move for non-monetary reasons, and so a person might not earn more income after migration (109). Surprisingly, dividing the population between the 25% with the highest and lowest income in 2016 does not yield such a strong division as much as age or gender (Figure 5). The average migration rates are $\lambda^{(l)} = 0.44$ for low income individuals and $\lambda^{(h)} = 0.42$ for high income individuals and profiles for both groups are similar. A more significant difference is observed in the return rate since the low-income population is more likely to return to previously-lived cities.

Workers in bigger cities are usually more productive and earn more than workers in smaller cities and rural areas (110; 111) although large cities disproportionately attract both high- and low-skilled workers (112). Here, we observe - perhaps counter-intuitively - that (at a country level) people with high and low income have similar migration patterns in terms of their rates and the signatures generated when moving, meaning that a person is almost equally likely to move within a country and to return if they have a low or a high income. Next, we investigate these age and gender divides at a city level, comparing large wealthy cities in the north of Colombia to mid-size and typically poorer cities in the south.

4.6 People from large cities move less and return more frequently

Beyond migration profiles for different population groups (such as women, men, young or mature), signatures also enable us to compare migration profiles between distinct cities. For a city (say A), we compute signatures similarly by looking at the frequency at which a person stays from one year to the next one ($AA\dots$), migrates ($AB\dots$) and returns ($ABA\dots$) and construct the *migration profile* of the city as follows. The *probability of moving*, μ_A, is the ratio between the number of migrations and the number of years in which locations are known. The *probability of returning* ρ_A is the frequency at which people who moved from city A return to live in it. Although the probability of moving and returning at a city level μ_A and ρ_A are similar to the migration and return rates at an individual level λ_i and π_i, but at a city level, the parameters indicate frequencies, whereas at an individual level they are rates and are estimated based on an unlabelled network via signatures. Results show that people from large cities, such as Bogotá, Medellín and Cali, have a small probability of moving. Also, if a person moves from these cities, there is a very high chance that the person will return (Figure 7). In smaller cities, the opposite happens. There is a 25% chance that a person from Buga, for example, will move, and only 1 in 8 of those who move will return to that city. As has been seen before (85), there

Figure here

Figure 7: Probability of moving μ (horizontal axis) and probability of returning ρ (vertical axis) for the 62 metropolitan areas and the countryside in Colombia (top panel). The bottom panel shows the same points, but in this case, our analysis considers four population groups (men, women, young and mature) for a subset of the largest cities separately. The coloured arrows indicate the probability of moving μ and returning ρ for men (yellow), women (orange), the young (blue) and the mature (red) for each of these cities.

Figure here

Figure 8: Age divide per city where the height of each bar (left) or the size of the disc (right) represents the city size. Longer bars correspond to a larger divide, and the national level of age and gender divide is represented by the vertical dark lines.

are nonlinear effects between city size and migration, and people from larger cities are much less likely to move and more likely to return than their small-city counterparts.

Next, we apply our methodology to different population groups *within* the largest cities (bottom panel of Figure 7). The gender and the age divide observed at a national level is also observed at a similar scale at a city level. Overall, we find that women and mature populations tend to move less, while young and male groups tend to move more. A young person from the countryside, for example, is 2.4 times more likely to move than a mature person and 7% less likely to return.

Thus far, we have identified significant age and gender differences in migration signatures. Yet, this marked difference is not homogeneously observed across the country. In some cities, men and women have similar migration and return rates, whereas in other parts, particularly small cities and the countryside, the gender - and age - divide is more pronounced. This type of city-level disparity is the mechanism through which accelerated ageing and gender imbalance in a city is deepened in some parts of the country (10; 8). We quantify the *age divide* for each city as the euclidean distance between the probability of moving μ and the probability of returning ρ between the mature and the young population groups. The *gender divide* is quantified similarly as the distance between the probability of moving μ and returning ρ between women and men. A larger divide, whether it is age or gender, means more distinct migration patterns between the population groups.

We find that the age and gender divide are not homogeneous across Colombia (Figure 8). In Bogotá and Cali, there is a large age divide, indicating that young people from those two cities move more frequently, but both cities have a smaller gender divide than the national average. The opposite happens in Medellín and Barranquilla, where the gender divide is much larger than the age divide. In fact, Medellín has the smallest age divide. In general terms, cities from south Colombia have a greater age divide than a gender divide, whereas, in northern Colombia, the age divide and gender divide are of similar magnitude. In the north part of Colombia, many large and industrialised cities attract nearby workers, whereas, in the rural parts of the south, the age and gender divide is much more substantial than in most parts of the country.

4.7 People leave the countryside and rarely return

Roughly one in three people from Colombia live in the countryside, and due to the nature of our data, it is impossible to trace movements within. In order to tackle this issue, we have merged all municipalities that are not part of Colombian metropolitan areas into a compacted region called *countryside*. Therefore, if a person moved, for instance, from Punta Gallinas (Cape Hens), the northernmost point on the mainland of South America, to Leticia, the southernmost point of Colombia in the Amazon river, 1865 kilometres away, we ignore that internal migration. We find that the 'countryside' unit is one of the zones with the highest probability of moving (to a metropolitan area) and a very small probability of returning (to any place within the countryside). In particular, the youngest individuals are very likely to move to a metropolitan area and not return (see Figure 7, where the countryside is represented by a single observation). Furthermore, the countryside is the observation with the highest age and gender divide (Figure 8). A young person in the countryside is nearly 2.5 times more likely to move than a mature person.

Some countries observe a rapid urbanisation process from rural to urban migration (5). Although Colombia is already highly urbanised, with more than 80% of its population being urban, we still observe a high flow of formal employees from the countryside to its cities. There are two significant implications of this process. Firstly, considering that only formal employees are observed in our data, the fact that they tend to move out and not return highlights the challenge of creating formal and steady jobs outside the country's core cities. Skilled people and people who gain knowledge and experience from work move too frequently out of the countryside, making the process of creating formal jobs even more challenging. Secondly, the fact that young individuals are nearly three times more likely to move than mature people accelerates the ageing process of the countryside. The observed selective migration process aggravates the existing gaps between large metropolitan areas and the countryside.

5 Discussion

Internal migration is critical for urban dynamics. With the progressive convergence of birth and death rates between countries, migration is the principal source of population re-distribution within countries (4). Migration is a specific type of human mobility characterised by a lower frequency and much more extended periods of stay. These features pose a severe challenge to most data-driven mobility models as individuals need to be observed for many years to detect aspects such as repeat and return patterns. Furthermore, most data sets span a limited time window, leaving out migrations that occurred before the recorded period. We propose a generative algorithm that aims to reproduce migration patterns or signatures based on population-level parameters of onwards and return migration. Our method is based on estimating the speed of migrating and returning, so it overcomes the starting point issue. Furthermore, our results summarise a complex process using only two parameters: the migration rate λ and the return rate π.

Migration can be observed through many different perspectives and techniques, depending on the research question and the available data. From the standpoint of computational social science, we observe a general pattern in how society moves. Observing migration through administrative data and a model offers a variety of insights that would not be captured with census data, particularly in terms of repeat and return migration.

Although rare, we observed some people who moved seven or even eight times in nine years and, in some cases, most of their migrations are returns. Access to an administrative dataset enabled us to analyse nine years of location data for individuals who work in the formal sector in Colombia and describe their migration patterns. However, far from being a representative sample of the country, it focuses on a specific set of people: formal employees who worked (almost) continuously between 2008 and 2016. Hence, we have a biased population sample in terms of age, income, gender and other demographics, so we cannot observe retirement or student migration, for example. Part of the gaps observed (e.g., between genders) is due to this bias. Also, besides the population bias, we capture the city where a person lives based on the address of their job, which might not correspond to the actual location, particularly under remote working schemes. Although still limited in most countries, administrative records provide valuable inputs to understand and forecast demographic trends faster and more precisely than census data.

Our results highlight a very substantial challenge in terms of gender. We observe that women experience a wage gap (113), are less likely to be formal employees (40% of the formal market in 2016 were women), and are also less likely to pass our filtering procedure (38.6% of the filtered observations are women). This indicates that women are more likely than men to work only for a few years in the formal economy or to work intermittently. Women are more likely to stay in the same city; women who move do it at a lower rate than their male counterparts, tend to return to previously-lived cities, and travel a shorter distance. The gender divide is larger than the age divide. Most of these differences are likely to be rooted in the country's gender inequality, with substantial implications for Colombia's mobility and productivity.

Between one year and the next, only 5.5% of people will move and from those who move, 19.4% will return one year later. Thus, looking only at three consecutive years of data, it would be concluded that migration and returning are very infrequent. However, our novel estimation method generates opposing results. Roughly 30% of people will eventually move, and some people move many times, so migration is highly concentrated among a few individuals (as it happens with wealth). Most human activities tend to concentrate in different ways. For example, a small population group suffers most of the crimes (114), but also a small group commits the majority of the crimes (115). Many commercial and social activities are highly concentrated. For instance, the richest person has more wealth than the population of various countries combined. Considering the migration rate λ_i as the variable of interest, we compute the Gini index across all individuals i in a standard manner (see the Supplementary Information for the corresponding Lorentz curves). We find that the top 5% migratory individuals accumulate nearly 25% of the migration rate in Colombia and the top 24% accumulate 80% of the migration rate. Similar to the dichotomy observed for daily mobility patterns (54), here we observe a large population group that will never move and a small population that moves with high intensity. The Gini index of the migration rate is high, with a coefficient of 0.7206 for all Colombia, but it is even more concentrated for mature (0.7942) and female (0.7501) populations.

Return migration is very common. Three out of four times that a person moves, after their first migration, they will move back to previously lived cities. Thus, a large part of the internal migration flow is going back. However, there is a significant difference between the young and mature, between men and women and between cities. Young people from the countryside are highly likely to move and not return, as opposed to mature people from large cities, who are likely not to move, and if they do, they almost

certainly return.

Finally, high levels of outward migration are concerning for secondary cities in Colombia. Smaller cities tend to experience a larger outflow of people with fewer returns compared to Medellín, Cali and Bogotá. These cities create less formal jobs and have a lower formality rate (93). Thus, small secondary cities are lagging metropolitan areas, less capable of creating formal employment, and are less likely to keep their workers or attract them to return after their first migration and compete with primary cities (116). The challenge is substantial as 68% of the population in Colombia lives in one of the 62 metropolitan areas considered here, with 43% residing in a city with more than one million inhabitants and 57% in smaller cities. Most Colombians live in a secondary city with less than one million inhabitants (25.2%), where people are more inclined to leave and never return. Our results indicate that small cities in Colombia will undergo an accelerated ageing process due to internal migration. The issue is even more complex for the countryside, home to 31.2% of the total population but less than 7% of formal employees. Young people from the countryside, particularly males, are very likely to move to a city and not return to the countryside, worsening the gender balance and accelerating the ageing of its population. A similar process is also expected in other parts of the world with a high intensity of inter-urban migration (14), and so it is likely that small cities in Mexico or Brazil, for example, will experience a similar accelerated ageing process.

6 Data availability

We use administrative records of the social security system in Colombia (abbreviated as PILA in Spanish, meaning the Integrated Report of Social Security Contributions), which contains job information about all formal workers in Colombia. We are unable to directly share the raw data. However, there is a protocol for gaining secure access to the data via the Ministry of Finance and Public Credit (MFPC) of Colombia or the Ministry of Health and Social Protection (MHSP) of Colombia. We followed that protocol and gained permission to use the PILA dataset for the current study. Please contact the ministries directly for more information.

7 Code availability

No relevant code was produced for this manuscript.

References

[1] Khalid Koser. Why migration matters. *Current History*, 108(717):147, 2009.

[2] Hermanus Geyer. Expanding the theoretical foundation of differential urbanization. *Tijdschrift voor Economische en Sociale Geografie*, 87(1):44–59, 1996.

[3] June JH Lee, Jill Helke, and Frank Laczko. *World Migration Report 2015*. International Organization for Migration, Geneva, Switzerland, 2015.

[4] Martin Bell, Elin Charles-Edwards, Philipp Ueffing, John Stillwell, Marek Kupiszewski, and Dorota Kupiszewska. Internal migration and development: Com-

paring migration intensities around the world. *Population and Development Review*, 41(1):33–58, 2015.

[5] Philip Rees, Martin Bell, Marek Kupiszewski, Dorota Kupiszewska, Philipp Ueffing, Aude Bernard, Elin Charles-Edwards, and John Stillwell. The impact of internal migration on population redistribution: An international comparison. *Population, Space and Place*, 23(6):e2036, 2017.

[6] Michael J Greenwood. Internal migration in developed countries. *Handbook of Population and Family Economics*, 1:647–720, 1997.

[7] Vincent Verbavatz and Marc Barthelemy. The growth equation of cities. *Nature*, 587(7834):397–401, 2020.

[8] Steffen Kröhnert and Sebastian Vollmer. Gender-specific migration from Eastern to Western Germany: Where have all the young women gone? *International Migration*, 50(5):95–112, 2012.

[9] Hannah Ritchie and Max Roser. Gender ratio. *Our World in Data*, 2019.

[10] Amina Maharjan, Siegfried Bauer, and Beatrice Knerr. Do rural women who stay behind benefit from male out-migration? a case study in the hills of Nepal. *Gender, Technology and Development*, 16(1):95–123, 2012.

[11] David Carr. Population and deforestation: why rural migration matters. *Progress in Human Geography*, 33(3):355–378, 2009.

[12] Jorge Rodríguez-Vignoli and Francisco Rowe. How is internal migration reshaping metropolitan populations in Latin America? a new method and new evidence. *Population Studies*, 72(2):253–273, 2018.

[13] Jorge De la Roca. Selection in initial and return migration: Evidence from moves across Spanish cities. *Journal of Urban Economics*, 100:33–53, 2017.

[14] Wilbur Zelinsky. The hypothesis of the mobility transition. *Geographical Review*, pages 219–249, 1971.

[15] Adam Dennett. *Understanding internal migration in Britain at the start of the 21st century*. University of Leeds, 2010.

[16] Aude Bernard, Francisco Rowe, Martin Bell, Philipp Ueffing, and Elin Charles-Edwards. Comparing internal migration across the countries of Latin America: A multidimensional approach. *PloS one*, 12(3):e0173895, 2017.

[17] Robert McGinnis. A stochastic model of social mobility. *American Sociological Review*, pages 712–722, 1968.

[18] George C Myers, Robert McGinnis, and George Masnick. The duration of residence approach to a dynamic stochastic model of internal migration: a test of the axiom of cumulative inertia. *Eugenics Quarterly*, 14(2):121–126, 1967.

[19] Amelie F. Constant and Klaus F. Zimmermann. The dynamics of repeat migration: A Markov chain analysis. *International Migration Review*, 46(2):362–388, 2012.

[20] Katherine Stovel and Marc Bolan. Residential trajectories: The use of sequence analysis in the study of residential mobility. *Sociological Methods & Research*, 32(4):559–598, 2004.

[21] Tom Kleinepier, Helga AG de Valk, and Ruben van Gaalen. Life paths of migrants: A sequence analysis of Polish migrants family life trajectories. *European Journal of Population*, 31(2):155–179, 2015.

[22] Jonathan Zufferey, Ilka Steiner, and Didier Ruedin. The many forms of multiple migrations: Evidence from a sequence analysis in Switzerland, 1998 to 2008. *International Migration Review*, 55(1):254–279, 2021.

[23] Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007.

[24] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[25] United Nations High Commissioner for Refugees UNHCR PROGRAMME. Global report, 2019, 2019.

[26] Moritz UG Kraemer, Adam Sadilek, Qian Zhang, Nahema A Marchal, Gaurav Tuli, Emily L Cohn, Yulin Hswen, T Alex Perkins, David L Smith, Robert C Reiner, et al. Mapping global variation in human mobility. *Nature Human Behaviour*, 4(8):800–810, 2020.

[27] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, and Marcello Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018.

[28] Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

[29] Ernest George Ravenstein. The laws of migration. *Journal of the Statistical Society of London*, 48(2):167–235, 1885.

[30] Samuel A Stouffer. Intervening opportunities: a theory relating mobility and distance. *American Sociological Review*, 5(6):845–867, 1940.

[31] Morton Schneider. Gravity models and trip distribution theory. *Papers in Regional Science*, 5(1):51–56, 1959.

[32] Joshua J Lewer and Hendrik Van den Berg. A gravity model of immigration. *Economics Letters*, 99(1):164–167, 2008.

[33] Donata Gnisci. West african mobility and migration policies of OECD countries. *Paris, Organisation for Economic Cooperation and Development, Sahel and West Africa Club*, 2008.

[34] WFP. At the root of exodus: Food security, conflict and international migration. World Food Programme, 5 2017.

[35] Wim Naude. *Conflict, disasters and no jobs: Reasons for international migration from Sub-Saharan Africa*. Number 85 in 1. UNU-WIDER, 2008.

[36] Monica Alexander, Kivan Polimis, and Emilio Zagheni. Combining social media and survey data to nowcast migrant stocks in the United States. *Population Research and Policy Review*, pages 1–28, 2020.

[37] K Bruce Newbold. Primary, return and onward migration in the US and Canada: Is there a difference? *Papers in Regional Science*, 76(2):175–198, 1997.

[38] K Bruce Newbold. Counting migrants and migrations: Comparing lifetime and fixed-interval return and onward migration. *Economic Geography*, 77(1):23–40, 2001.

[39] K Bruce Newbold and Martin Bell. Return and onwards migration in Canada and Australia: Evidence from fixed interval data. *International Migration Review*, 35(4):1157–1184, 2001.

[40] Andrea Cattaneo and Sherman Robinson. Multiple moves and return migration within developing countries: A comparative analysis. *Population, Space and Place*, 26(7):e2335, 2020.

[41] Aude Bernard, Martin Bell, and Yu Zhu. Migration in China: A cohort approach to understanding past and future trends. *Population, Space and Place*, 25(6):e2234, 2019.

[42] Aude Bernard and Francisco Perales. Is migration a learned behavior? understanding the impact of past migration on future migration. *Population and Development Review*, 2021.

[43] Jane Falkingham, Jo Sage, Juliet Stone, and Athina Vlachantoni. Residential mobility across the life course: Continuity and change across three cohorts in Britain. *Advances in Life Course Research*, 30:111–123, 2016.

[44] Michel Poulain, Anne Herm, and Roger Depledge. Central population registers as a source of demographic statistics in Europe. *Population*, 68(2):183–212, 2013.

[45] Alina Sîrbu, Gennady Andrienko, Natalia Andrienko, Chiara Boldrini, Marco Conti, Fosca Giannotti, Riccardo Guidotti, Simone Bertoli, Jisu Kim, Cristina Ioana Muntean, et al. Human migration: the big data perspective. *International Journal of Data Science and Analytics*, pages 1–20, 2020.

[46] Moshe Levy. Scale-free human migration and the geography of social networks. *Physica A: Statistical Mechanics and its Applications*, 389(21):4913–4917, 2010.

[47] Lee Fiorio, Emilio Zagheni, Guy Abel, Johnathan Hill, Gabriel Pestre, Emmanuel Letouzé, and Jixuan Cai. Analyzing the effect of time in migration measurement using georeferenced digital trace data. *Demography*, 58(1):51–74, 2021.

[48] Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. A tale of many cities: universal patterns in human urban mobility. *PloS One*, 7(5):e37027, 2012.

[49] Riccardo Di Clemente, Miguel Luengo-Oroz, Matias Travizano, Sharon Xu, Bapu Vaitla, and Marta C González. Sequences of purchases in credit card data reveal lifestyles in urban populations. *Nature Communications*, 9, 2018.

[50] Marta C González, Cesar A Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[51] Renaud Lambiotte, Vincent D Blondel, Cristobald De Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, 2008.

[52] Shan Jiang, Joseph Ferreira, and Marta C Gonzalez. Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. *IEEE Transactions on Big Data*, 3(2):208–219, 2017.

[53] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, 2010.

[54] Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. Returners and explorers dichotomy in human mobility. *Nature Communications*, 6:8166 EP –, 09 2015.

[55] S. Rinzivillo, L. Gabrielli, M. Nanni, L. Pappalardo, D. Pedreschi, and F. Giannotti. The purpose of motion: Learning activities from individual mobility networks. In *2014 International Conference on Data Science and Advanced Analytics (DSAA)*, pages 312–318, Oct 2014.

[56] Luca Pappalardo, Maarten Vanhoof, Lorenzo Gabrielli, Zbigniew Smoreda, Dino Pedreschi, and Fosca Giannotti. An analytical framework to nowcast well-being using mobile phone data. *International Journal of Data Science and Analytics*, 2(1):75–92, Dec 2016.

[57] Christian M Schneider, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C González. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246, 2013.

[58] John Stillwell, Peter Boden, and Adam Dennett. Monitoring who moves where: Information systems for internal and international migration. In *Population Dynamics and Projection Methods*, pages 115–140. Springer, 2011.

[59] Julie DaVanzo. Repeat migration in the United States: who moves back and who moves on? *JStor*, 1980.

[60] Michael J Thomas, John CH Stillwell, and Myles I Gould. Modelling the duration of residence and plans for future residential relocation: A multilevel analysis. *Transactions of the Institute of British Geographers*, 41(3):297–312, 2016.

[61] Julie DaVanzo. Repeat migration, information costs, and location-specific capital. *Population and Environment*, 4(1):45–73, 1981.

[62] Adrian J Bailey. Getting on your bike: what difference does a migration history make? *Tijdschrift voor Economische en Sociale Geografie*, 80(5):312–317, 1989.

[63] Sidney Goldstein. The extent of repeated migration: An analysis based on the Danish population register. *Journal of the American Statistical Association*, 59(308):1121–1132, 1964.

[64] Guy Pourcher. Un essai d'analyse par cohorte de la mobilité géographique et professionnelle. *Population (French edition)*, pages 357–378, 1966.

[65] Sidney Goldstein. Repeated migration as a factor in high mobility rates. *American Sociological Review*, 19(5):536–541, 1954.

[66] Peter A Morrison. Chronic movers and the future redistribution of population: a longitudinal analysis. *Demography*, 8(2):171–184, 1971.

[67] Julie S DaVanzo and Peter A Morrison. Return and other sequences of migration in the United States. *Demography*, 18(1):85–101, 1981.

[68] Jaciel Arce Montoya, Renato Salas Alfaro, and José Antonio Soberón Mora. La migración de retorno desde Estados Unidos hacia el Estado de México: oportunidades y retos. *Cuadernos Geográficos*, 1(49):153–178, 2011.

[69] Larry Long. *Migration and residential mobility in the United States*. Russell Sage Foundation, New York, USA, 1988.

[70] David A Plane and Jason R Jurjevich. Ties that no longer bind? the patterns and repercussions of age-articulated migration. *The Professional Geographer*, 61(1):4–20, 2009.

[71] David A Plane, Christopher J Henrie, and Marc J Perry. Migration up and down the urban hierarchy and across the life course. *Proceedings of the National Academy of Sciences*, 102(43):15313–15318, 2005.

[72] Andrei Rogers. Age patterns of elderly migration: an international comparison. *Demography*, 25(3):355–370, 1988.

[73] David A Plane. Demographic influences on migration. *Regional Studies*, 27(4):375–383, 1993.

[74] Aude Bernard. Levels and patterns of internal migration in Europe: A cohort perspective. *Population Studies*, 71(3):293–311, 2017.

[75] Rory Coulter, Maarten van Ham, and Allan M Findlay. Re-thinking residential mobility: Linking lives through time and space. *Progress in Human Geography*, 40(3):352–374, 2016.

[76] Martin Bell. How often do Australians move? alternative measures of population mobility. *Journal of the Australian Population Association*, 13(2):101–124, 1996.

[77] Douglas S Massey and Rene M Zenteno. The dynamics of mass migration. *Proceedings of the National Academy of Sciences*, 96(9):5328–5335, 1999.

[78] Mariana Macedo, Laura Lotero, Alessio Cardillo, Hugo Barbosa, and Ronaldo Menezes. Gender patterns of human mobility in Colombia: Reexamining Ravensteins laws of migration. In *Complex Networks XI*, pages 269–281. Springer, 2020.

[79] Laetitia Gauvin, Michele Tizzoni, Simone Piaggesi, Andrew Young, Natalia Adler, Stefaan Verhulst, Leo Ferres, and Ciro Cattuto. Gender gaps in urban mobility. *Humanities and Social Sciences Communications*, 11(1), 2020.

[80] Philip S Morrison and William AV Clark. Loss aversion and duration of residence. *Demographic Research*, 35:1079–1100, 2016.

[81] Scott M Myers. Residential mobility as a way of life: Evidence of intergenerational similarities. *Journal of Marriage and the Family*, pages 871–880, 1999.

[82] Aude Bernard and Sergi Vidal. Does moving in childhood and adolescence affect residential mobility in adulthood? an analysis of long-term individual residential trajectories in 11 European countries. *Population, Space and Place*, 26(1):e2286, 2020.

[83] James E Anderson. The gravity model. Technical report, National Bureau of Economic Research, 2010.

[84] Aba Schwartz. Interpreting the effect of distance on migration. *Journal of Political Economy*, 81(5):1153–1169, 1973.

[85] Rafael Prieto Curiel, Luca Pappalardo, Lorenzo Gabrielli, and Steven Richard Bishop. Gravity and scaling laws of city to city migration. *PloS one*, 13(7):e0199892, 2018.

[86] Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

[87] Marc Barthélemy. *Spatial networks*. Springer, USA, 2014.

[88] Allen C Kelley and Leonard W Weiss. Markov processes and economic analysis: the case of migration. *Econometrica: Journal of Econometric Society*, pages 280–297, 1969.

[89] Neil W Henry, Robert McGinnis, and Heinrich W Tegtmeyer. A finite model of mobility. *The Journal of Mathematical Sociology*, 1(1):107–118, 1971.

[90] Adam Dennett. Estimating flows between geographical locations: get me started in spatial interaction modelling. *Technical report*, 2012.

[91] Mikaela Backman, Esteban Lopez, and Francisco Rowe. Career trajectories and outcomes of forced migrants in Sweden: Self-employment, employment or persistent inactivity? *Small Business Economics*, 2018.

[92] Gilles Duranton. Delineating metropolitan areas: Measuring spatial labour market networks through commuting patterns. In *The Economics of Interfirm Networks*, pages 107–133. Springer, 2015.

[93] Neave OClery, Rafael Prieto Curiel, and Eduardo Lora. Commuting times and the mobilisation of skills in emergent cities. *Applied Network Science*, 4(1):118, 2019.

[94] Jinzhou Cao, Qingquan Li, Wei Tu, Qili Gao, Rui Cao, and Chen Zhong. Resolving urban mobility networks from individual travel graphs using massive-scale mobile phone tracking data. *Cities*, 110:103077, 2021.

[95] Rafael Prieto Curiel, Sofía Collignon Delmar, and Steven Richard Bishop. Measuring the distribution of crime and its concentration. *Journal of Quantitative Criminology*, pages 1–29, 2017.

[96] Rafael Prieto Curiel and Steven Richard Bishop. A measure of the concentration of rare events. *Scientific Reports*, 6, 2016.

[97] Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[98] Dankmar Böhning, Ekkehart Dietz, and Peter Schlattmann. Recent developments in computer-assisted analysis of mixtures. *Biometrics*, pages 525–536, 1998.

[99] Peter Schlattmann, Johannes Hoehne, and Maryna Verba. *CAMAN: Finite Mixture Models and Meta-Analysis Tools - Based on C.A.MAN*, 2016. R package version 0.74.

[100] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[101] Andrei Rogers and Luis J Castro. Model migration schedules. *International Institute for Applied Systems Analysis (IIASA)*, 1981.

[102] Adam Dennett and John Stillwell. Internal migration in Britain, 2000–01, examined through an area classification framework. *Population, Space and Place*, 16(6):517–538, 2010.

[103] Adam Dennett and John Stillwell. Internal migration patterns by age and sex at the start of the 21st century. In *Technologies for Migration and Commuting Analysis: Spatial Interaction Data Applications*, pages 153–174. IGI Global, 2010.

[104] Michael P Todaro. A model of labor migration and urban unemployment in less developed countries. *The American Economic Review*, 59(1):138–148, 1969.

[105] Guido Dorigo and Waldo R Tobler. Push-pull migration laws. *Annals of the Association of American Geographers*, 73(1):1–17, 1983.

[106] Andres J Garcia, Deepa K Pindolia, Kenneth K Lopiano, and Andrew J Tatem. Modeling internal migration flows in Sub-Saharan Africa using census microdata. *Migration Studies*, 3(1):89–110, 2015.

[107] Everett S Lee. A theory of migration. *Demography*, 3(1):47–57, 1966.

[108] Xingna Nina Zhang, Wenfei Winnie Wang, Richard Harris, and George Leckie. Analysing inter-provincial urban migration flows in China: A new multilevel gravity model approach. *Migration Studies*, 07 2018.

[109] Larry A Sjaastad. The costs and returns of human migration. *Journal of Political Economy*, 70(5, Part 2):80–93, 1962.

[110] PP Combes, G Duranton, L Gobillon, and S Roux. Estimating agglomeration effects with history, geology, and wordker fixed effects. *The Economics of Agglomeration, Nationla Bureau of Economic Research, Cambridge, MA*, 2010.

[111] Jorge De La Roca and Diego Puga. Learning by working in big cities. *The Review of Economic Studies*, 84(1):106–142, 2017.

[112] Jan Eeckhout, Roberto Pinheiro, and Kurt Schmidheiny. Spatial sorting. *Journal of Political Economy*, 122(3):554–620, 2014.

[113] Alejandro Hoyos, Hugo Ñopo, and Ximena Peña. The persistent gender earnings gap in Colombia, 1994-2006. *Documento CEDE*, 1(2010-16), 2010.

[114] SooHyun O, Natalie N. Martinez, YongJei Lee, and John E. Eck. How concentrated is crime among victims? a systematic review from 1977 to 2014. *Crime Science*, 6(1):9, 2017.

[115] Natalie N. Martinez, YongJei Lee, John E. Eck, and SooHyun O. Ravenous wolves revisited: a systematic review of offending concentration. *Crime Science*, 6(1):10, Aug 2017.

[116] Brian Herbert Roberts and Rene Peter Hohmann. The system of secondary cities: The neglected drivers of urbanising economies. *CIVIS Sharing Knowledge and Learning from Cities*, 7, 2014.

8 Acknowledgements

Anonymised manuscript.

9 Author contributions

Anonymised manuscript.

10 Competing interests

Anonymised manuscript.

11 Materials and Correspondence

All correspondence should be addressed to
Anonymised manuscript.