

Using Clustering Algorithms to Characterise Uncertain Long-term Decarbonisation Pathways

Pei-Hao Li ^{a,*}, Steve Pye ^a, Ilkka Keppo ^a

^a UCL Energy Institute, University College London, Central House, 14 Upper Woburn Place, London, WC1H 0NN, UK

Abstract

Long-term decarbonisation pathways to achieve ambitious low-carbon targets involve a range of uncertainties. Different energy system modelling approaches can be used to systematically evaluate the influence of the uncertainties, but this often leads to an unmanageable number of pathways. Summarising the large ensemble through a more limited number of representative pathways, to inform stakeholders, can be challenging. This study thus explores how to identify representative decarbonisation pathways using clustering algorithms, which can assist in grouping similar data points in uncategorised datasets, such as pathway ensembles. However, the suitability of clustering algorithms for pathway characterisation has not been investigated to date. Hence, k-means, hierarchical clustering, Gaussian mixture model, spectral clustering, and density-based clustering are adopted for comparisons. An illustrative pathway ensemble for the United Kingdom is applied to evaluate their performance based on cluster validity indices. Three metric transformations, including power, standardisation and sectoral standardisation, are also applied to create three additional sets of pathways for testing. The k-means algorithm is found to outperform others consistently, although hierarchical clustering might also be applicable if the distribution of pathway proximity is uneven. The results also highlight the utility of the approach in revealing distinctive trade-offs between technologies among the identified representative pathways. For instance, the electrification of heating can be replaced by district heating in the residential sector. The described, novel approach can be applied to characterise other sets of pathways, with greater technological details generated by any energy system models, to reveal insights for long-term decarbonisation.

* Corresponding author. Tel: +44 20 31087801
E-mail address: p.li@ucl.ac.uk (P.H. Li)

Keywords: uncertainty; decarbonisation pathways; energy transition; clustering analysis

1. Introduction

Energy system models (ESMs) are an essential tool to help policy-makers and researchers investigate how to transform national energy systems, e.g. in order to achieve the ambitious targets set by Paris Agreement [1]. These ESMs, such as TIMES models [2], typically adopt a cost optimisation-based framework, exploring least-cost energy system configurations, subject to greenhouse gas (GHG) emissions targets and other constraints. To explore possible low-carbon technology portfolios, these models take into account numerous technologies in all energy-related sectors. For instance, the Irish TIMES model and TIMES-Sweden comprise about 1,700 and 1,800 technologies respectively [3,4].

Nonetheless, the projection of future technology costs and other characteristics (e.g. efficiency, lifetime) is highly uncertain. In addition to the continuous, but uncertain, technology improvements, unexpected technology breakthroughs, sudden policy changes, global demand variations, social acceptance issues etc. [5] can significantly alter the prospects of individual technologies. For example, before Tesla Motor entered the car market, it was hard to imagine the high acceptance of electric vehicles (EVs) [6], with no major car manufacturers being willing to offer battery EVs on a commercial basis in early 2010s [7]. With breakthroughs in fast charging technologies [8] and longer driving ranges [9], EV sales have increased to 2.6% in the Netherlands and to near 40% in Norway in their car markets in 2017 [10]. Traditional car manufacturers, consequently, showed their ambitious in the global EV market [11]. Moreover, in the UK, the introduction of feed-in tariff in 2010 has boosted the installation of small scale solar photovoltaic (i.e. less than 10 kW) [12], followed by a sudden plummet in 2016 due to a dramatic cut to feed-in tariff for solar, announced by the UK government [13]. In addition, future electricity demand has been shown to be significantly affected by future temperature and economy, both are highly uncertain [14]. Finally, even though smart grid technologies can increase system flexibility to accommodate intermittent renewable energy, the potential of smart grid technologies heavily depends on public's acceptance, which is influenced by perceived ease of use and risk, and is thus uncertain [15]. Consequently, the long-term decarbonisation pathways

suggested by ESMs, under varying assumptions about future technology performance, lead to significantly different low-carbon technology portfolios. While scenario analyses are a common and useful approach for investigating the influences of uncertain future assumptions, in practise they usually considers only a limited number of futures around a few key technologies or system features. For example, only five scenarios were considered for Denmark in [16], and four scenarios for China in [17]. Six global scenarios were generated considering various climate uncertainties and GHG reduction targets [18]. Panos et al. used two scenarios to investigate benefits of flexibility options for long-term decarbonisation of Switzerland [19]

In recent years, there have been several studies focusing on developing new approaches to systematically explore techno-economic uncertainties, in order to help policy-makers make more robust decisions. For example, global sensitivity analysis (GSA) is applied in Fais et al. [20] to explore the influences of uncertainty across a range of low-carbon technologies on the 28 generated long-term decarbonisation pathways. The Modelling to Generate Alternatives (MGA) approach has been used to explore the uncertainty of model structure by identifying possible near-optimal pathways [21]. This approach shows that the model structure, in respect of what is or is not explicitly modelled, together with the formulation of optimality, can also strongly affect the conclusions one draws about the role different technologies could play in the future. For instance, Price and Keppo [22] used MGA to reveal elements that are robust across global decarbonisation pathways, even when differences to previous runs are explicitly rewarded in the objective function. Similarly, MGA was applied to explore UK's potential decarbonisation pathways that might deviate from the cost optimal ones [23]. In addition, Monte-Carlo analysis (MCA) based approaches have also been adopted to systematically investigate the influences of multiple uncertain input parameters, represented by probability distributions, on long-term decarbonisation pathways [24]. A large number of pathways can be fairly easily generated with this technique. For example, Pye et al. [25] generated 500 decarbonisation pathways and Li and Trutnevyte [23] created 800 different scenario pathways. Moreover, 1800 and 2048 decarbonisation pathways were created for the UK [26] and for the world [27] respectively using MCA to reflect the uncertainties in the energy system transtions. For more detailed review on these approaches, please refer to Yue et al. [24]. Using these

systematic approaches, especially MCA, to explore uncertain decarbonisation pathways comprehensively can, therefore, often generate large amounts of data, e.g. tens of thousands of technology metrics, due to the numerous technologies in ESMs.

However, approaches for helping decision makers determine a limited number of distinct strategic alternatives (as reflected by the representative pathways) amongst the numerous plausible long-term decarbonisation pathways are still lacking to date. In most of previous studies, only simple data visualisation measures, correlation analysis or regression models were applied to explore the most influential technologies or resources on predefined system-wide indicators, such as GHG emissions and total system costs. For example, Fais et al. [20] used box-and-whisker plots and correlation coefficients to investigate the impacts of a few key technological changes on energy consumption and GHG emissions across 28 scenarios. In Li and Trutnevyte [23], visualisation of installation capacity of electricity generation technologies, such as fossil-fuel, nuclear power plants, and renewable energy, across 800 scenarios were applied to show the variation in technology deployment levels. Only maximally-different pathways were further investigated, instead of trying to identify ones that would best represent the full sample. Pye et al. [25], in turn, focused on realising the influences of a few key uncertain factors, such as biomass availability, gas prices, and nuclear capital costs, on total GHG emissions across 1000 pathways using scatter plots and linear regressions between predefined variables. Again, scatter plots were adopted by Fragkos et al. [27] to explore the relationships between crucial system indicators, such as carbon prices and gas reserves, across 2048 scenarios. No representative decarbonisation pathways were revealed from their scenarios and the indicators used to illustrate the pathway ensemble were handpicked by the modellers. Similarly, visualisation of the electricity mix by technology across 200 pathways was created to quantify wind deployment capacity in 2050 for Denmark [28]. More recently, clustering algorithms have also been applied to investigate uncertain pathways, but still focused on realising the influence of individual technologies. For example, Moksnes et al. [29] used Gaussian mixture model (GMM) to group similar scenarios based on estimated system costs and GHG emissions alone. Determinants of grouped scenarios were then identified by the Patient Rule Induction Method. Pye et al. [26] applied a hierarchical clustering (HC) algorithm to group highly correlated technologies across 1800 decarbonisation pathways,

but focusing on the interdependencies between individual technologies, as opposed to full pathways. Overall, past studies have largely used ad hoc approaches for illustrating differences between the pathways and the focus has been on illustrating uncertainty in outcomes, rather than trying to summarise the large number of pathways as a smaller set of distinct strategic alternatives. Finally, the few studies that have used more advanced, less user dependent, methods have not focused much on the exact formulation of the problem or the choice of the algorithms to use.

Identifying a few representative pathways from a huge pathway ensemble can be challenging. This is due to the large, multidimensional information embedded in pathways, including all technology metrics over the modelling years. Moreover, the absence of label/category on generated pathways makes the task even more daunting. A systematic approach is thus desirable to characterise these long-term pathways, to extract what might be the key types of pathways that together represent the larger ensemble.

Clustering algorithms, a type of unsupervised machine learning approach, could be useful in characterising a long term energy system pathway ensemble, something that they have not been applied for previously. These algorithms are used for grouping unlabelled, multi-dimensional data points (in our case pathways, as reflected through their pathway metrics) based on their characteristics so that data points in a cluster are as similar to each other, and as distinct from those in other groups, as possible [30]. Such algorithms could thus provide a promising approach for defining the distinct pathway families, or clusters, that differ in terms of the strategies they suggest for the development of the future energy system.

Nevertheless, different clustering algorithms group data points based on various presumptions on the distribution of the target dataset. The performance of these algorithms is thus extremely sensitive to the actual distribution of data points. Given the structure of the pathway dataset is not known a priori, it is thus essential to try and assess various clustering algorithms to determine an appropriate one to characterise decarbonisation pathways, as different algorithms often result in entirely different partitions even on the same data [31]. In this paper, we explore the application of clustering algorithms to the analysis of decarbonisation scenarios, a field to which this approach has been rarely applied in the past, and assess the performance of different techniques.

This study thus aims to fill the research gap by applying five different clustering algorithms to characterise long-term decarbonisation pathways.

The contribution of this study is threefold:

1. Explore how to apply clustering algorithms to investigate a large number of detailed, long-term decarbonisation pathways
2. Compare the suitability and relative strengths of a range of clustering algorithms for this task
3. Identify distinct decarbonisation pathway families to summarise the “archetype alternatives” for the highly uncertain system transition toward a low-carbon target

This paper is structured as follows: applications of clustering algorithms in energy-related fields are firstly reviewed in Section 2. The research approach for assessing the application of five different clustering algorithms is briefly introduced in Section 3. Then, Section 4 compares the performances of the five clustering algorithms, based on five cluster validity indices, and discusses the identified representative pathways. Finally, some conclusions on the application of clustering algorithms for finding representative pathways are drawn in Section 5.

2. Application of clustering algorithms for energy related studies

Clustering algorithms have been widely adopted in a range of energy studies, covering the evaluation of energy performance of buildings [32], identification of representative demand patterns of buildings [33], improvement of prediction ability for short-term energy demand of buildings [34], identification of level of energy efficiency of industrial sub-sectors [35], design of distributed generators [36], wind resource analysis and prediction [37], among others.

In the field of energy and buildings, Schütz et al. [33] used several clustering algorithms, such as k-means (KM) and k-medoids, to cluster energy demands of two buildings into representative periods to reduce the computational complexity of energy system optimisation models for buildings. They found that k-medoids performed the best in terms of designing an energy system close to the optimal one. Papadopoulos et al. [38] applied KM to group commercial and housing buildings in New York City into clusters with distinct

energy use intensities over a six-year period, to explore the influential factors on the temporal patterns of energy performance for these identified representative buildings. Larger, newer, and higher valued commercial buildings were found to be the main driver of energy reduction in the building sector. Unternährer et al. [36] used KM to cluster buildings according to their spatial locations to simplify the required input data for an optimisation model used to identify potential areas for deploying district heating. Tang et al. [34] grouped energy consumption records of a building with a KM algorithm for energy consumption prediction. The prediction errors of the built models based on clustered records reduced considerably. Yilmaz et al. [39] applied KM to identify representative daily electricity demand profiles of Swiss households using five features of daily profiles (mean of morning, daytime, evening, night and standard deviation). Three typical daily demand profiles were found for households. Four distinct demand profiles were further revealed for a single household to represent the variety of demand profiles over time. The findings suggest that a mean demand profile cannot represent households' demand characteristics well. Boudet et al. [40] used KM to reveal households' typical energy-saving behaviours using five energy-saving attributes, including cost, frequency of performance, required skill level, observability, and locus of decision. Four distinct energy-saving behaviours were found.

Clustering techniques have also been applied for other end-use sectors. For instance, Liao and He [35] adopted a clustering algorithm to categorise the levels of energy efficiency of 37 industrial sub-sectors in China, to evaluate the influencing factors on their energy performance. Technological progress, energy consumption structure, and enterprise scale were found to be the most influential factors. Arias and Bae [41] used a HC algorithm to classify traffic patterns based on real-world traffic volume data in South Korea. A decision tree-based model was then developed to classify traffic pattern types, considering a set of influential factors, such as maximum temperature and average humidity, for the prediction of EV charging demand.

Furthermore, clustering algorithms have been proven to be a valuable technique in wind power analysis and prediction. For example, Dong et al. [42] used KM to divide meteorological data into clusters, comprising records with similar characteristics. Another regression model was then applied to reveal the patterns in the records in a cluster with

similar meteorological characteristics to the prediction day. Using this approach, the accuracy of day-ahead wind prediction was improved considerably. Similarly, Sun et al. [43] used a hybrid-clustering method to group similar wind speed days, which have similar wind variance trends, into clusters as training data to build a prediction model for short-term wind power forecasts (hours ahead). The prediction accuracy for two wind farms in China was improved with this model.

However, clustering algorithms have rarely been applied to the broader energy system – or to characterise energy system wide transition pathways. To the best knowledge of the authors, Csereklyei et al. [44] is the only example, applying a model-based clustering algorithm to categorise the energy mix in 28 EU member states over past 40 years into 7 groups. This was done to investigate the major drivers on national energy transitions from high to low fossil fuel systems over time. Lower shares of fossil fuels were found to be typically associated with higher national income and energy use per capita. However, as this research used the energy ladders concept, it focused only on primary energy use. ESM-generated decarbonisation scenarios, which include large amounts of information on the potential roles of individual technologies, have not been used with clustering approaches to date.

3. Methodology

3.1 Research procedure

Our approach to the research, illustrated in **Figure 1**, has been developed to explore how clustering algorithms can be used to explore representative pathways from large scenario ensembles. First, we selected our ensemble of scenarios, representing decarbonisation pathways generated based on multiple techno-economic uncertainties and taken directly from a previous study [26]. The various pathway metrics (e.g. use of specific fuel for heating, or electricity produced with a specific technology) were then subjected to three different transformation approaches, including power-transformation, standardisation and sectoral standardisation. This was done so that the considered algorithms can be tested with different ensembles of pathways with various distribution of technology metrics to determine the most robust one for characterising pathways. Proximity matrices were then

constructed using the original and the transformed pathways. Five clustering algorithms were then used to analyse the original as well as the transformed ensembles of pathways for comparison and exploration of the suitability for grouping specific decarbonisation pathways as reflected through their technology metrics (in 2030 and 2050). Each algorithm was chosen from one of five major clustering algorithm categories, including distance-based, model-based, density-based, agglomerative and graph-based clustering algorithms. Algorithms from different categories are likely to generate different results [31,45]. The performance of clustering algorithms was then compared based on several performance indices to reveal the most suitable one for the clustering analysis done for our specific dataset. Finally, five groups of pathways were determined using the best performing algorithm. Centroid pathways, which were comprised of the mean metrics of those of pathways in individual clusters, were regarded as representative pathways and further discussed.

Scikit-learn [46], a widely-adopted and well-developed machine learning toolkit, was used to carry out the clustering analyses.

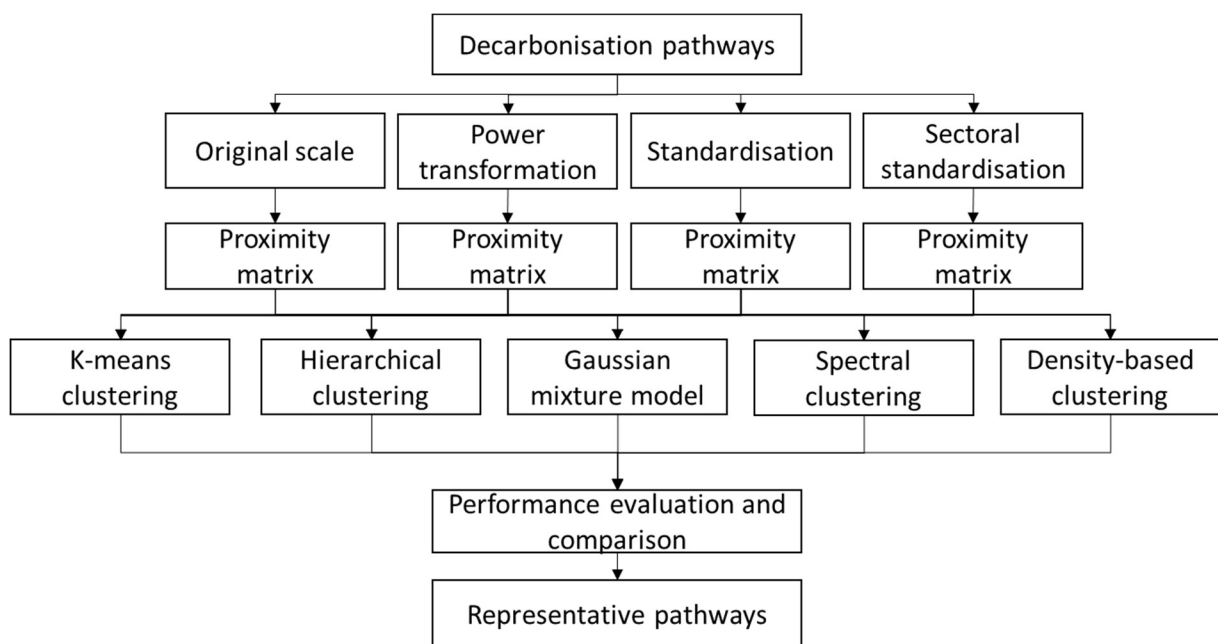


Figure 1. Research procedure to characterise decarbonisation pathways

3.2 Decarbonisation pathways under techno-economic uncertainties

A set of 600 decarbonisation pathways, all achieving UK’s 2050 GHG emissions targets, were taken from Pye et al. [26] for demonstrating our approach. These pathways were

generated with the Energy System Modelling Environment model (ESME) [47], a technology-explicit whole energy systems model for the UK, by Monte Carlo sampling of techno-economic parameters [25] of the model. ESME uses linear programming to determine portfolios of technologies across all sectors of the UK energy system that lead to the lowest cumulative discounted total energy system cost until 2050, while simultaneously satisfying the UK decarbonisation targets of 53% reduction in GHG emissions by 2030 and 80% by 2050. In the parameter sampling, not only the uncertainties of technology costs and resource costs were taken into account, but also the uncertainties related to technology build rates and resource availability, such as maximum potential of biomass produced sustainably in the UK or via import from the international market. The variance of parameters was set according to the maturity of technologies, with a higher variance for novel or emerging technologies, for which the uncertainties are still greater. Probability functions were constructed to represent the uncertainty of the considered parameters using triangular distributions [26]. For each simulation, values for intermediate years (prior to 2050) were determined based on interpolation back to the base year (2010) value, following a linear trajectory between 2010 and 2050. For more details regarding pathway generation under techno-economic uncertainties, please refer to Pye et al. [26].

3.3 Proximity matrix

A proximity matrix, representing relationship between any two pathways, were constructed for clustering algorithms to group similar pathways. The proximity (distance) between pathways is defined as follows:

$$d_{i,j} = \sqrt{\sum_y \sum_{m=1}^M (x_{i,y,m} - x_{j,y,m})^2} \quad (1)$$

where $x_{i,y,m}$, $x_{j,y,m}$ are the metrics of technology m in pathway i and j in year y ; M is the total number of technologies in a pathway. In this study, only technology metrics in 2030 and 2050 were considered for simplification. A lower proximity value means the two pathways under consideration are more similar to each other than pathways with a higher value. A proximity matrix was then created to describe the distance between any two pathways. The size of the full matrix is thus $N \times N$, where N is the total number of pathways. This matrix can then be used as the basis for the clustering analysis.

3.4 Metric transformation

As clustering algorithms are sensitive to the distribution of data points, it is essential to test those algorithms with different distributions of pathways to determine the most robust one. Three metric transformation approaches, including power transformation, standardisation, and sectoral standardisation, are thus applied to twist the distribution of the adopted pathways. Four ensembles of pathways were then applied for sensitivity analysis of the performance of each clustering algorithm.

In addition, the metric transformations can help to reduce the dominance of a few, high volume metrics, so that the importance of other metrics in the end-use sectors can also be stressed in the clustering processes. Otherwise qualitatively different decarbonisation strategies may be missed by the algorithm, as even small relative changes for a dominating variable could dwarf qualitatively more dramatic shifts in other variables. Consequently, metric transformations provide additional flexibility for the pathway grouping, by focusing on different perspectives of technology switching across pathways. Several transformation approaches are tested, as while specific approaches can mitigate certain problems, they can also introduce new ones.

The three proposed transformation approaches are explained as follows.

i) Power transformation:

Power transformation is used to transform technology metrics using a power function with a coefficient of 0.35 to lower the scale of high metric values:

$$x'_{i,y,m} = x_{i,y,m}^{0.35} \quad (2)$$

where $x'_{i,y,m}$ is the transformed metric m in pathway i in year y .

The coefficient 0.35 is adopted to bring the transformed value close to that one would get from using natural logarithm transformation. However, unlike natural logarithm transformation, power transformation will not cause numerical difficulties when a metric has zero value.

As the high variance of metrics is significantly reduced, the distribution of the transformed pathways is more condensed than that of the original pathways. The variance

of end-use metrics becomes more evident, compared to that of supply-side metrics, after transformation.

ii) Standardisation:

With standardisation, the metrics of pathways are transformed with average metrics and metrics' standard deviations. The transformation can be formulated as follows.

$$x'_{i,y,m} = \frac{x_{i,y,m} - \bar{x}_{y,m}}{std_{y,m}} \quad (3)$$

where $\bar{x}_{y,m}$ and $std_{y,m}$ is the mean and standard deviation of technology metric m across pathways in year y respectively. As a result, every metric has an influence more closely linked to relative changes, even those with extremely low absolute values. Thus while this approach reduces the dominance of high absolute value variables, the transformation could lead to an exceptionally irregular distribution of pathways.

iii) Sectoral standardisation:

Sectoral standardisation transformation, in turn, replaces the standard deviation of individual metrics with the average standard deviation of all metrics in the sector to which the metrics belong.

$$x'_{i,y,m} = \frac{x_{i,y,m} - \bar{x}_{y,m}}{std_{y,s}} \quad (3)$$

where $std_{y,s}$ is the sectoral standard deviation of the sector s across pathways in year y . Metrics will have a stronger influence when their variance is relatively higher than that of the other metrics in the same sector. This transformation can thus emphasise the technology replacement *within* a sector, by equalising the relative weights of the sectors. However, the variance of metrics with extremely low absolute values would not be dramatically amplified as that transformed with standardisation.

3.5 Clustering algorithms

Five clustering algorithms are adopted for the comparison: KM, HC, GMM, Spectral clustering (SC), and Density-based spatial clustering of applications with noise (DBSCAN), all of which are explained briefly below.

3.5.1 K-means (KM)

KM is one of the most popular clustering algorithms to group similar data points into a predefined C number of clusters. The algorithm initialises C centroids randomly from all the data points. The rest of the data points are then assigned to their nearest centroids to form C clusters. Centroids, in turn, are updated as the means of the data points in clusters. The assigning of data points and updating of centroids are repeated iteratively until the centroids remain the same [30].

This algorithm performs well when the data points are gathered together in circular-like groups [30]. In addition, due to its clustering procedure, the grouped clusters tend to have similar numbers of data points [48].

3.5.2 Hierarchical clustering (HC)

HC groups data points based on the proximity between data points through an iterative procedure and ultimately generates a hierarchy of clusters [30]. Both bottom-up (agglomerative) and top-down (divisive) approaches can be adopted to determine the clusters. The agglomerative approach starts with treating each data point as an individual cluster. The most similar pairs of clusters are then merged to form larger clusters iteratively until there is a single cluster that includes all data points. The divisive approach, on the other hands, starts with a single cluster containing all data points and divides clusters into smaller ones until each cluster contains only one data point. The order of cluster merging determined by the HC algorithm can be represented as a tree-like diagram, dendrogram, with leaves correspond to clusters. A specific number of clusters can then be determined by intersecting the tree through a proximity chosen by the algorithm. In this study, the widely used bottom-up approach is adopted to iteratively merge a pair of clusters in which the most dissimilar members are closest, known as complete-linkage merge strategy.

This algorithm is typically used to understand the underlying hierarchical structure of data points. Furthermore, there is no need to choose an anticipated number of clusters beforehand. However, once two subsets of points are merged at an intermediate step, the

grouped arrangement cannot be changed later on, potentially leading to a locally optimal solution. Due to this, the approach may not be suitable for noisy and high-dimensional data [30].

3.5.3 Gaussian Mixture Model (GMM)

A GMM is a probabilistic model that can fit a mixture of a finite number of multivariate Gaussian distributions to the distribution of data points by adjusting the shapes of the probability distributions [49]. Each subset of data points that can be represented by a single or multiple Gaussian distributions forms a cluster. As such, the clustering problem transforms to a statistical learning problem, to optimise the number of model components (Gaussian distributions), associated weights, and the parameters of each component [50].

An expectation-maximisation (EM) algorithm is applied to determine the unknown parameters of the Gaussian distributions, such as the means and covariances. The EM algorithm is an iterative procedure with two alternating stages, expectation (E) and maximisation (M) stages. In the E stage, parameters are first estimated by maximising the log-likelihood function of parameters and then, based on the Bayes rule, each data point can be assigned to a specific cluster with the highest probability for the specific point. Then, in the M stage, the parameters are derived by maximising the log-likelihood function again under new posterior probabilities. The E and M stages are repeated iteratively until the improvement of the log-likelihood is smaller than a stopping criterion [49,51].

Compared with KM, GMM is more flexible as it can group data points distributed in any kind of elliptical shapes and assigns a data point to multiple clusters probabilistically [44,52].

3.5.4 Spectral clustering (SC)

SC algorithm performs dimensionality reduction before clustering data points [49,53]. Data points are transformed into a weighted, undirected graph from the proximity matrix, representing the distances between data points. When the constraints on the assignments of points to clustering are relaxed from binary conditions to continuous status, the data clustering problem is then transformed into a spectral partition problem. This can be dealt with eigenanalysis techniques, cutting the set of points into separate subsets of points, where the similarity among points in the same subset is high and across different subsets is low. The algorithm can be summarised as follows [54]:

Step 1: Construct a proximity matrix for the sample data.

Step 2: Calculate the first k eigenvalues and eigenvectors to construct an eigenvector space.

Step 3: Cluster the eigenvectors within the eigenvector space with a clustering algorithm, such as KM.

In this study, KM is adopted to determine the clustering of data points; therefore, a predefined number of clusters is required as an input parameter.

Unlike KM and GMM, SC can be applied to sample spaces of any shape and perform clustering analysis on a non-convex distribution, a polygon with one or more of interior angles is more than 180 degrees [49,54]. However, this algorithm performs poorly to divide datasets that contains structures at different scales of size and density [55].

3.5.5 Density-based spatial clustering of applications with noise (DBSCAN)

DBSCAN is also a popular algorithm for grouping data points into clusters, with the advantage of not requiring as input a predefined number of clusters [30]. Points in high-density regions are aggregated together and separated by low-density regions. Those points in low-density regions are treated as noise and ignored, removing the influence of outliers on the clustering.

This algorithm classifies all points into three categories, including core, border, and noise points, according to two user-defined parameters, radius (Eps) and minimum number points ($MinPts$). Core points are those with at least $MinPts$ neighbouring points within the distance of Eps , which is relatively small. Border points are not core points but within the neighbourhood of core points. The rest of the points are noise points and are discarded. Edges are put between core points if they are within a distance Eps of each other. Core points connected by edges are grouped into a separate cluster. Border points are then assigned to the nearest clusters.

Consequently, this algorithm is relatively resistant to noise (outliers) and can handle data points distributed in arbitrary shapes and sizes. However, it might have trouble with high-dimensional data or data distributed with widely varying densities [30].

3.6 Performance evaluation and the choice of number of clusters

Since there are no labels indicating the assignments of pathways to individuals clusters, it is challenging for clustering algorithms to determine the optimal number of clusters [31]. Instead, cluster validity indices are usually applied to evaluate the performance of clustering algorithms and to provide a sensible way to determine a suitable number of clusters. These indices measure the performance of a clustering structure based on cluster cohesion and cluster separation, the former indicating how closely related data points in a cluster are, and the latter how well-separated the clusters are from each other [30]. As no one single index can adequately reflect clustering performance, several indices, explained as follows, are adopted for performance comparisons of clustering algorithms and for informing a suitable choice for the number of clusters.

In the following sections, arrows next to index names in headings suggest the direction of value change for better performance of a clustering algorithm.

3.6.1 Sum of the squared error (↓)

Sum of the squared error (SSE) is a measure to represent the cohesion of clusters. It is simply the summation of squared difference between individual data points (pathways) in clusters and their centroids (i.e. the mean of the metrics of pathways for each cluster) and can be evaluated as follows [30]:

$$SSE = \sum_{c=1}^C \sum_{i \in c} \sum_y \sum_{m=1}^M \|x_{i,y,m} - \bar{x}_{c,y,m}\|^2 \quad (4)$$

where $\bar{x}_{c,y,m}$ is the mean technology metric m of pathways in the cluster c in year y ; C is the total number of clusters. A lower coefficient indicates better performance of a clustering algorithm as the data points in clusters are closer to each other.

3.6.2 Davies-Bouldin index (↓)

The Davies-Bouldin (DB) index estimates cluster cohesion based on the average distance from the points in a cluster to its centroid and cluster separation based on the distance between cluster centroids. The index is defined as follows:

$$DB = \frac{1}{C} \sum_{c=1}^C \max_{c'} \left\{ \frac{S_c + S_{c'}}{\|\bar{x}_c - \bar{x}_{c'}\|} \right\} \quad (5)$$

where c' is another cluster other than cluster c ; \bar{x}_c is the centroid of cluster c ; S_c is the average distance of data points of the cluster c to its centroid. As cluster separation is used as the denominator, a better clustering structure would have a lower value of the index.

3.6.3 Calinski-Harabaz index (\uparrow)

The Calinski-Harabaz (CH) index is another widely adopted measure for unlabelled clustering analysis, with a higher value suggesting better performance of a clustering algorithm. Cluster cohesion is estimated based on the distances from the pathways in a cluster to its centroid. Cluster separation, on the other hand, is based on the distance from clusters' centroids to the global centroid. The index can be defined as follows:

$$CH = \frac{\sum_{c=1}^C n_c \|\bar{x}_c - \bar{x}_g\|^2 / (C-1)}{\sum_{c=1}^C \sum_{i \in c} \|x_i - \bar{x}_c\|^2 / (N-C)} \quad (6)$$

where \bar{x}_c is the centroid of cluster c , which, as before, is simply the average of all pathways in the cluster; \bar{x}_g is the global centroid, which is the average of the centroids of all clusters; n_c is the number of pathways in cluster c [30,56].

3.6.4 Dunn index (\uparrow)

The Dunn (DN) index, in turn, evaluates the performance of clustering algorithms with the minimal distance between points of different clusters (cluster separation) and the largest within-cluster distance (cluster cohesion) [56]. The index is simply the ratio of the separation to the cohesion. The index can be defined as follows:

$$DN = \frac{\min_{i \in c, j \in c'} \|x_i - x_j\|}{\max_{i, j \in c} \|x_i - x_j\|} \quad (7)$$

Consequently, a higher value of the index indicates a better performance of a clustering algorithm.

3.6.5 Silhouette index (\uparrow)

The Silhouette (SL) coefficient evaluates the performance of clustering algorithms with the relationship between the cluster cohesion (a) and cluster separation (b). Cluster cohesion is measured using the average distance between a data point and the other data points in the same cluster. Cluster separation, on the other hand, is the average distance

between a data point and all data points in the nearest cluster. The Silhouette coefficient then can be calculated as follows:

$$SL = \frac{1}{N} \sum_{c=1}^C \sum_{i \in c} \frac{b_i - a_i}{\max\{b_i, a_i\}} \quad (8)$$

where $\max\{b_i, a_i\}$ is the greater one of a data point's cohesion and separation value. When the Silhouette coefficient is higher, implying clusters are more condense and the distances between clusters are larger, the performance of the clustering analysis is regarded better [30,56].

3.6.6 Ideal number of clusters

These indices are often plotted against various numbers of clusters, to give an understanding of how the performance changes as a function of the number of clusters.

Elbow measure is usually applied to illustrate the improvement of SSE as the number of clusters increases. When drawing out the relationship between the SSE and the number of clusters, an elbow point, or turning point, of the line can be found. The number of clusters at this point is a natural choice for clustering, as the marginal improvement of performance is much lower after that point.

The other 4 indices are also plotted against various numbers of clusters, to judge the best choice for how many clusters to use. For indices where higher values indicate better performance, the corresponding number of clusters at the peak point could be a reasonable choice. On the other hand, for those where lower values suggest better performance, the number of clusters with the lowest index value in the diagram should be identified. However, the plotting approach does not always work as clusters might be considerably more intertwined or overlapping than those which can be separated clearly [30].

4 Results and Discussion

In this section, the characteristics of the adopted pathways are first discussed, followed by the exploration of the influences of the different metric transformations on the pathway clustering analysis. Then, the performance of clustering algorithms is evaluated based on the identified groups of transformed pathways and using the five cluster validity indices. Finally, five representative pathways, which are the centroid pathways reflecting the mean metrics of the pathways within individual clusters, are identified using the most robust

clustering algorithm. Centroid pathways are regarded as the representative pathways in this study, since grouped pathways are more likely to be similar to the centroid pathway of their cluster than that of other clusters. It needs to be noted, however, that the centroid pathways may well differ to some extent from all individual pathways within the cluster¹⁸.

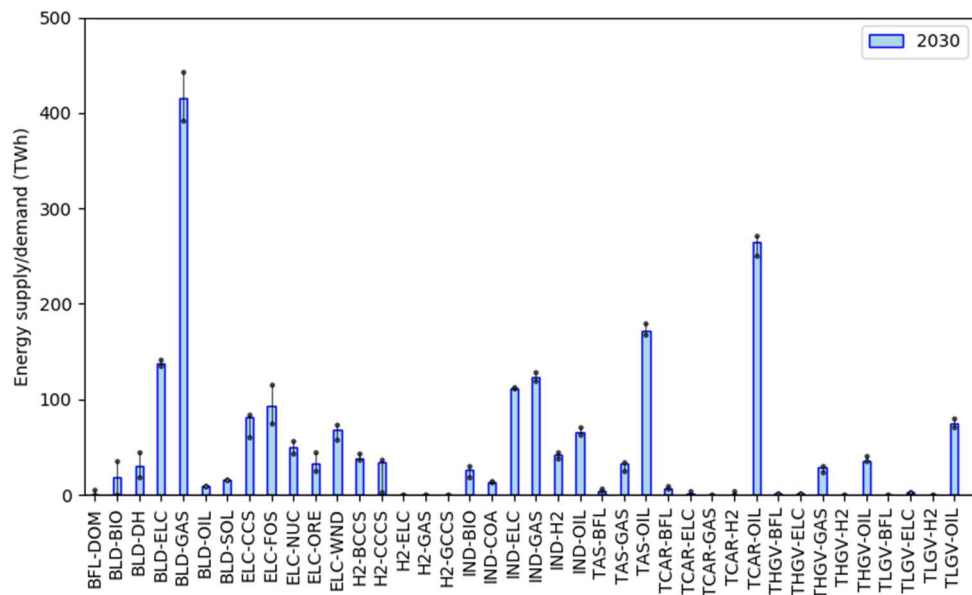
4.1 Set of scenarios for exploring clustering algorithms

Figure 2 shows the distributions of 40 aggregated technology metrics in 2030 and 2050 across the 600 pathways, creating a dataset of 48,000 values. Aggregated technology metrics were applied in this study for simplification. The metrics reflect sector and fuel specific activities and the definitions of the individual metrics can be found in Appendix A.

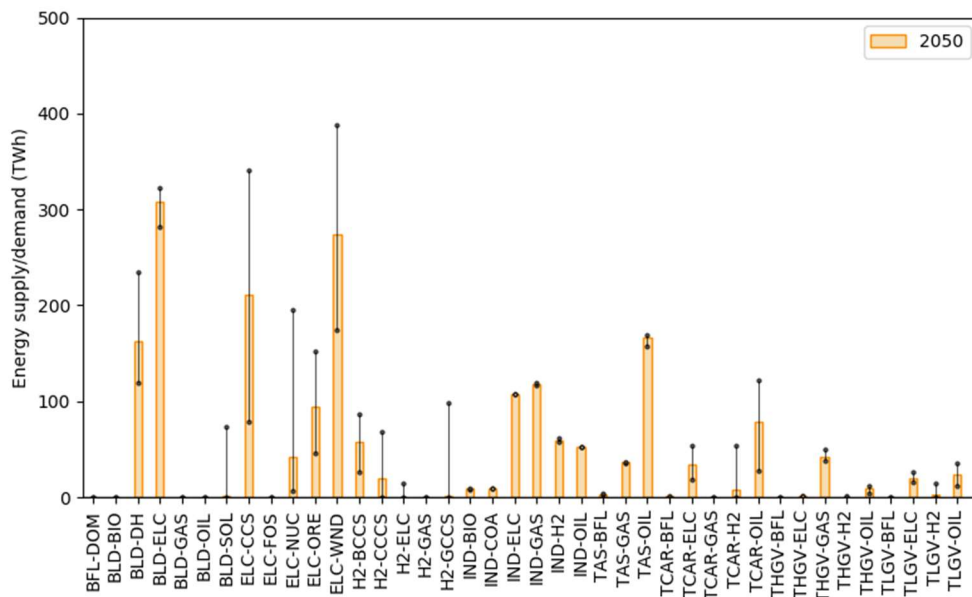
There is a clear trend of transitioning from more fossil-fuel based technologies in 2030 to low-carbon technologies in 2050 across all sectors. For example, in the building sector (BLD), the high gas consumption (BLD-GAS) for heating, hot water and cooking in 2030 is replaced by higher consumption of electricity (BLD-ELC) and use of more district heating for heat provision (BLD-DH). In the electricity sector, fossil fuel-based electricity generation (ELC-FOS) is fully phased out by 2050 and replaced by other low-carbon generation technologies, such as nuclear (ELC-NUC), wind (ELC-WIND), and power plants with carbon capture and storage (ELC-CCS), in 2050. Similarly, in the transport sector, oil consumption (TCAR-OIL) is reduced dramatically by 2050, with higher deployment of vehicles using electricity (TCAR-ELC). As for the industrial sector and the aviation sector, these two sectors see less mitigation of CO₂ as fewer technologies are available for decarbonisation. However, fuel-switching can still be found in these two sectors, e.g. more hydrogen is used to replace oil and gas in the industrial sector.

Some technology metrics, however, are highly uncertain across pathways, due to the uncertainty of future technology development, deployment speed limitations, etc. Obviously, the variance of technology metrics is much higher in 2050 than in 2030 as the technology mix in 2030 will more likely be influenced by the status of the existing energy system. Some technology metrics are more sensitive to the changes in technology assumptions. This leads to higher variances of those technologies in 2050, such as district heating, the majority of electricity generation technologies, hydrogen production

technologies, and oil-based vehicles. The high variability in resulting pathways due to technology uncertainty makes it difficult to identify representative pathways.



(a)



(b)

Figure 2 Distribution of technology metrics with narrower variations in (a) 2030 and wider variations in (b) 2050 across all pathways (bars show the average values, while vertical black lines indicate the maximum and minimum values of metrics in pathways).

4.2 Transformation approaches applied to scenarios

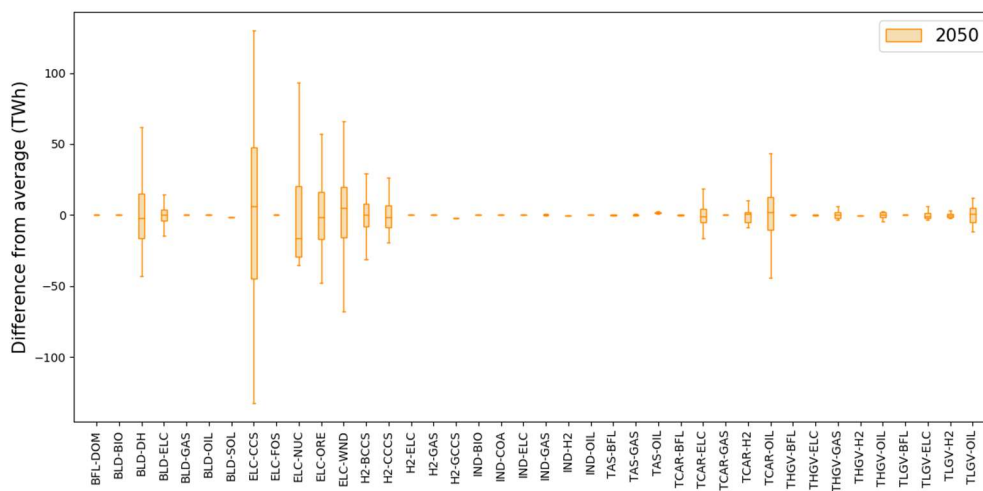
As clustering analysis is sensitive to distribution of proximities between pathways, three transformation approaches were applied to transform pathways for the comparison of the clustering algorithms. The effects of the different metric transformations are illustrated in **Figure 3**. As the proximity matrices are later formulated based on these distributions, the transformed distribution will influence how pathways are grouped by clustering algorithms. The influences of the transformation on clustering results will be addressed in the next section.

Figure 3(a) represents the distribution of differences between metrics and their averages. In the figure, the boxes represent ranges of the first quartile (Q1) and the third quartile (Q3) of the metric differences; whiskers indicate the possible highest and lowest values of metrics that are at most 1.5 times of interquartile range (Q3 – Q1) higher than Q3 and lower than Q1 respectively. Outliers beyond this range are not included. The variations are similar to those in **Figure 2**, except outlier cases, such as high solar thermal in buildings (BLD-SOL) and hydrogen production by gas (steam methane reformer, SMR) with CCS (H2-GCCS) in 2050 in **Figure 2**, which are outside the ranges covered by the box-and-whisker plots.

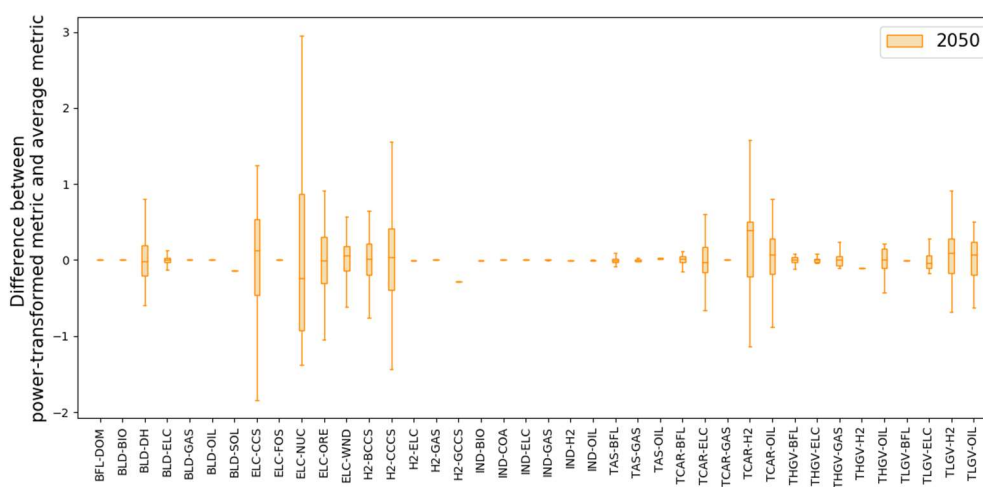
After the power-transformation, the scale of difference between transformed metrics and averages has been reduced dramatically, as shown in **Figure 3(b)**. Relatively speaking, the distribution of metrics with lower values was increased, such as hydrogen production by coal gasification with CCS (H2-CCCS) in 2050, whereas the distribution of metrics with higher values was reduced, such as electricity generation with CCS (ELC-CCS) in 2050. Those metrics with limited deviations remained insignificant, such as in the industrial sector.

As illustrated in **Figure 3(c)**, distributions of standardised differences between metrics and their averages became more similar. As this transformation aims to focus on the relative changes for a given metric, the variances of individual metrics are valued more equally than in the previous two transformation cases. This represents a case in which the possible dominance of large absolute value metrics is not only mitigated, but completely ignored, and thus very small absolute changes to marginal variables can change the metrics dramatically from one scenario to the other.

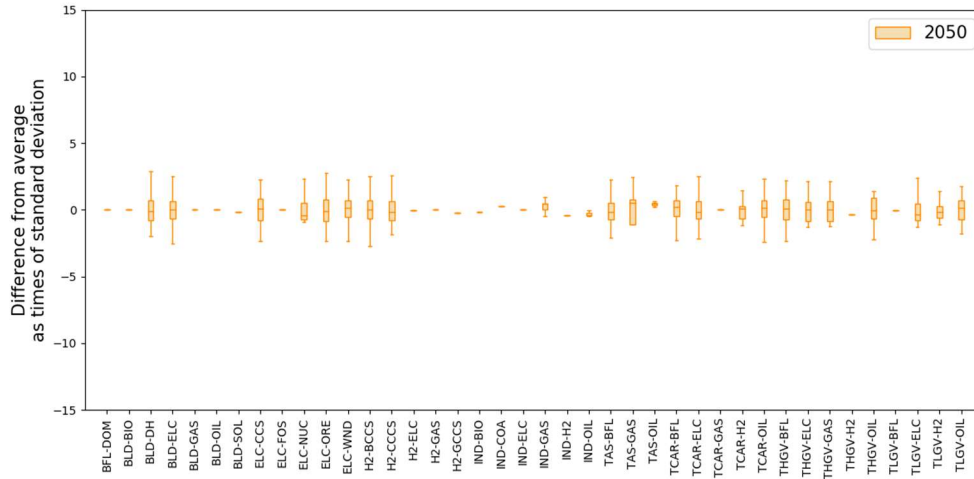
Finally, sectoral standardisation amplified the variances of metrics that are, compared to other metrics in the same sector, more uncertain (**Figure 3(d)**). For example, the variance of the oil consumption for light-good vehicles (TLGV-OIL) in 2050 was widened after transformation as the original variances of other transport technologies in 2050 are much smaller, as shown in **Figure 3(a)**. Conversely, the variance of electricity generation with CCS (ELC-CCS) in 2050 was lowered, as the variances of electricity technologies are generally high, again as shown in **Figure 3(a)**. This transformation thus aims to mitigate both, the dominating impact of high value metrics in a given sector and the impact of focusing only on relative changes. It thus stresses the importance of metric-switching within a sector, while mitigating the influences of metric replacement across sectors.



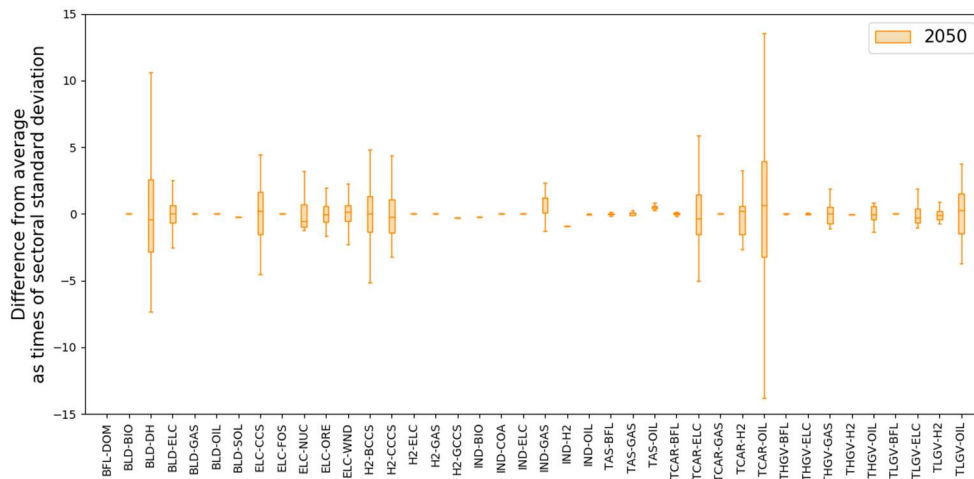
(a)



(b)



(c)



(d)

Figure 3 Distribution of differences between (a) individual metric and average metric, showing unproportionally wide variations of a few dominating metrics; (b) power-transformed metric and average metric, with reduced variations of the dominating metrics; (c) individual metric and average metric as times of standard deviation, with almost equal variations of all metrics; (d) individual metric and average metric as times of sectoral standard deviation, reducing variations of the dominating metrics while emphasising the importance of metrics in each sector, in 2050. Note the different scales of the y-axis.

The influences of metric transformation on representative pathways, the centroid pathways in pathway clusters determined by KM, are summarised in **Table 1**, with additional detail in Appendix C. Only KM is considered here for simplification. The major effects of the transformations on the grouped clusters are rescaling the boundaries of extreme metrics and shifting the signal strengths of metrics. For example, the extreme values of original metrics are in the range of 100 to -100 TWh (i.e. 360 to -360 petajoules),

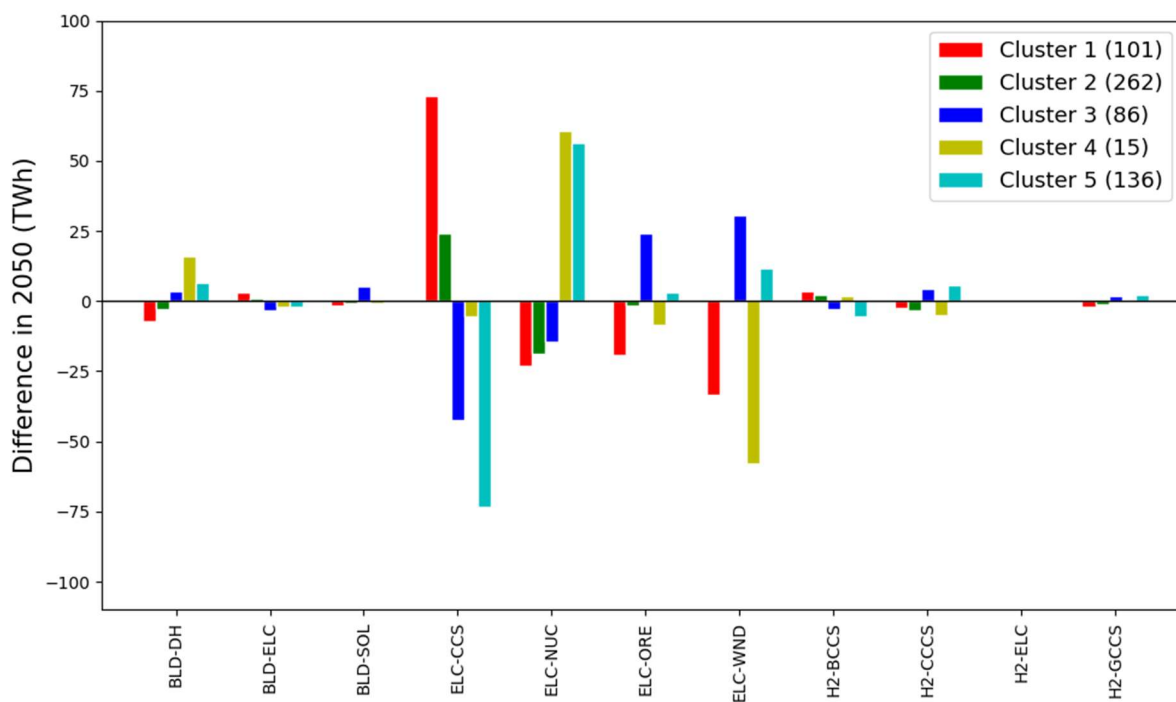
as shown in **Figure C.1**, whereas those values are much lower (between 60 and -60 TWh) (i.e. 216 and -216 petajoules) after applying power-transformation, as shown in **Figure C.2**.

Consequently, with transformed metrics the clustering analysis can alleviate the dominance of a few technology metrics in the power sector, so that the influence of some end-use metrics can be stressed in various ways. This is caused directly by the modified distribution produced by the transformation, as shown in **Figure 3**. For example, as shown in **Figure 4(b)** and **Figure 5(b)**, variance of oil cars (TCAR-OIL) and EVs (TCAR-ELC) in the clustered pathways based on sectoral standardised metric is more evident than that in the clustered pathways based on original metrics, while the variance of metrics in the power sector decreases. In **Figure 3(d)**, similar changes to the variations of those metrics can also be found. This allows one to observe clear, qualitative differences between the clusters, which may remain hidden, if the clustering is dominated by the few, high absolute value variables.

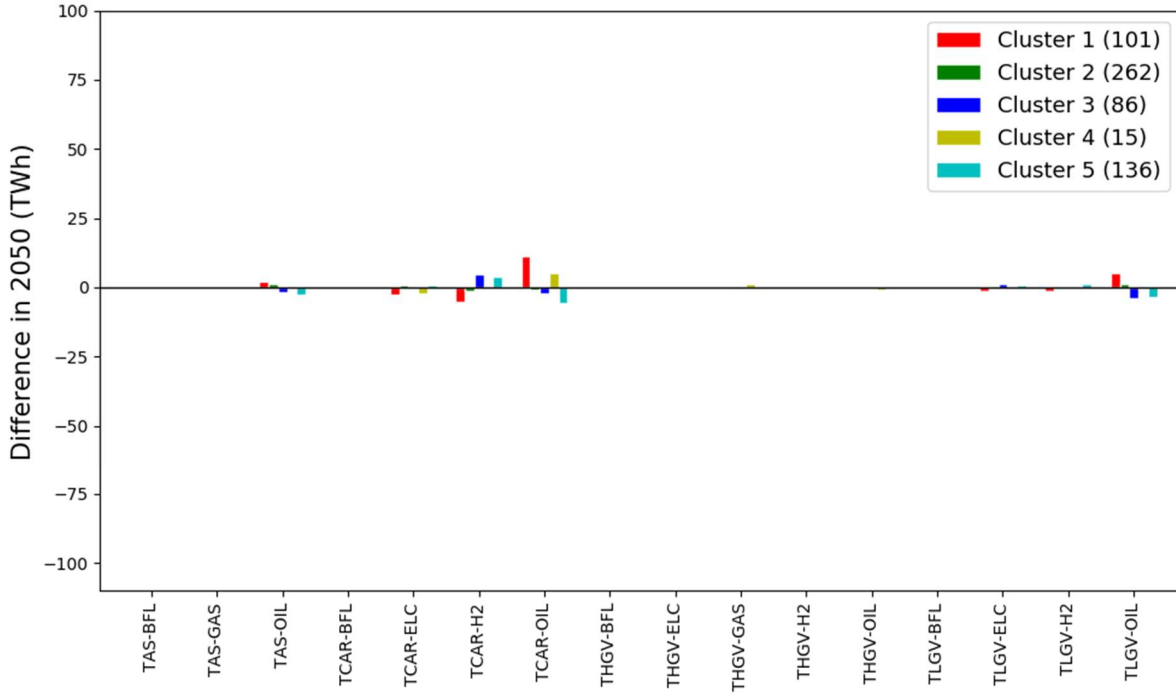
In other word, the metric transformation can emphasise the trade-offs between metrics across pathways from different perspectives. For instance, power transformation can stress the metric replacement across sectors as the dominance of a few metrics is mitigated. Sectoral standardisation, on the other hand, can highlight the trade-offs between metrics within individual sectors since the transformation highlights relationships between metrics within sectors.

Table 1 Influences of metric transformation on clustered metrics by k-means

Metric transformation	Characteristics of clustered metrics
None	<ul style="list-style-type: none"> All metrics are close to averages in 2030 Power sector is the only sector with obvious variation in 2050, as shown in Figure 4 and Figure C.1
Power	<ul style="list-style-type: none"> Influence of high variance of a few metrics, such as those in the power sector, is mitigated Variation of some metrics becomes more obvious, such as bioenergy consumption in the buildings (BLD-BIO) in 2030 and hydrogen production by gas and CCS (H2-GCCS) in 2050, as shown in Figure C.2
Standardisation	<ul style="list-style-type: none"> Variance of every metric is treated equally Variance of many clustered metrics is more obvious in both 2030 and 2050, as shown in Figure C.3
Sectoral standardisation	<ul style="list-style-type: none"> Relatively high variance of metrics in a sector are more likely to be revealed Trade-off between metrics in a sector is clearer, such as oil cars (TCAR-OIL) and EVs (TCAR-ELC) in 2050, as shown in Figure 5 and Figure C.4

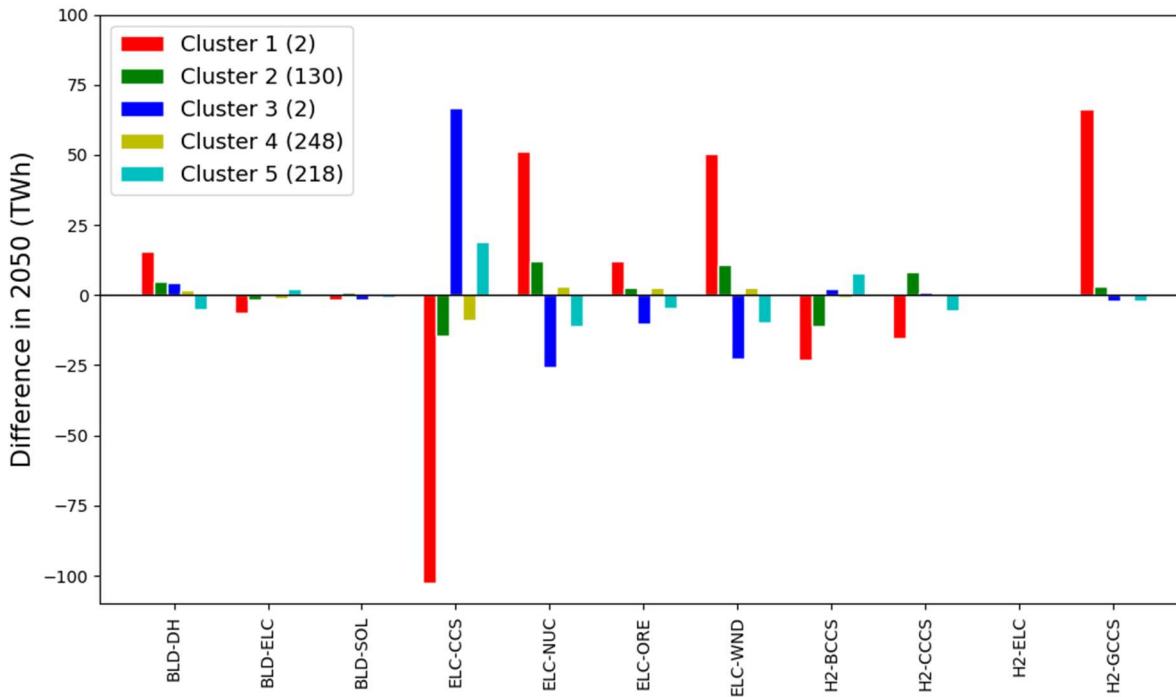


(a)

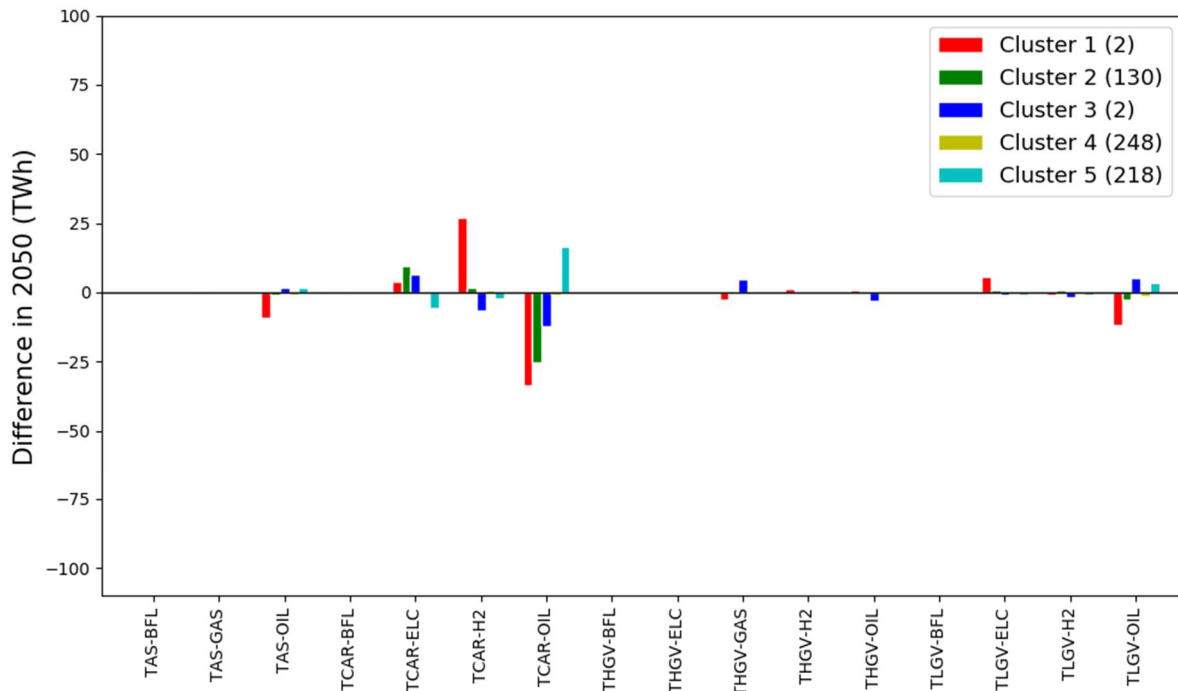


(b)

Figure 4 Differences between the average metrics of all pathways and the mean metrics of the five clusters in 2050 as identified by k-means based on original, untransformed metrics, showing that the clustering results majorly reflect the variations of a few key metrics in the electricity sector. Metrics of the building, electricity, and hydrogen sectors are shown in (a) and metrics of the transport sector are shown in (b).



(a)



(b)

Figure 5 Differences between the average metrics of all pathways and the mean metrics of the five clusters in 2050 as identified by k-means based on sectoral standardised metrics, showing that variations of metrics in both the electricity sector and the end-use sectors, such as the residential and the transport sectors, are emphasised. Metrics of the building, electricity, and hydrogen sectors are shown in (a) and metrics of the transport sector are shown in (b).

4.3 Comparison of clustering algorithms

Having transformed the data, each of the five clustering algorithm was applied to the original and the three transformed datasets, and assessed using a set of validity indices. The performance rankings of the clustering algorithms are listed in **Table 2**, with lower values indicating better performance. For example, based on the SSE index for the non-transformed dataset, KM performs best, whereas SC is the worst performing algorithm of the ones used. In **Figure 6**, the rankings based on various indices are aggregated to show the performance tendency of each individual algorithm. For more detailed comparisons, please refer to Appendix B. DBSCAN is not included in the table and figures as the number of clusters is determined automatically by the algorithm depending on the two essential input parameters.

Overall, KM performs the best based on the considered validity indices, e.g. the algorithm is ranked the best for 4 of the 5 indices, with original and power-transformed

pathways. In fact, KM scores highest in 65% of all comparisons. HC also performed well as it scores the best or the second best in 60% of comparisons. For GMM, it ranks second best in 40% of comparisons. It is obvious that SC is the worst performing algorithm as it is lowest ranked 95% of times. This implies that SC might not be a good choice for pathway groupings.

The better performance of KM, compared to GMM, may suggest the pathways are distributed continuously and densely in the feature space, but that the shape of those areas is probably fairly irregular. This would make it difficult for GMM to cover well with the same number of clusters as KM. The shape, in turn, could result from the ESME probability distribution functions of technology costs being triangular, leading to a higher probability for costs around the central value. Therefore, the joint distribution of cost assumptions of technologies would also be more likely to have a shape concentrating at a central area and spreading outward continuously as the number of data points drops. The resulting pathways were thus likely to be distributed closely in groups, separated as a result of the system dynamics from the ESM's sensitivity to technology cost changes. Similarly, since ESME is a linear model and can thus demonstrate abrupt changes to pathways resulting even from small input changes, known as "bang-bang" effect [57], this can also contribute to the irregular boundary of the pathway distribution.

As HC divided pathways into groups with a chosen minimum required distance between groups, distinct subsets of pathways, that are apart from each other less than the chosen proximity, might be more likely to be integrated into a larger group. KM, on the other hand, would not require a minimum distance between groups. A subset of such distinct pathways thus has a higher chance to be grouped as a standalone cluster. Therefore, KM outperforms HC in most of our cases. HC does, however, outperform KM when standardised pathways are clustered, given that HC is ranked as the best in 3 of the 5 indices, as shown in **Table 2**. This is even more evident when the number of cluster is low, as shown in **Figure B.3**. One possible explanation is that the standardisation made the distribution of pathways extremely uneven in the feature space as the variances of all technologies were rescaled to a similar range, as illustrated in **Figure 3(c)**, meaning that any technology, even if insignificant in its absolute value contribution, could easily influence the proximity between pathways.

The worst performing algorithm, SC, might suffer from the distribution of the pathways in the feature space. This algorithm represents the distribution of pathways as a graph in an eigenspace and tries to determine superior splitting, but the dense and continuous distribution of pathways means the vertices (pathways) in the transformed graph could still be closely connected without clear separations. This could make it difficult for the algorithm to find proper separations, resulting in lower indices scores. The poor clustering with SC for condense datasets has also been found by Nadler and Galun [55].

While not presented here, it is evident that DBSCAN is not suitable for grouping pathways. No matter the transformation used, DBSCAN always grouped a high percentage of pathways into a single cluster (e.g. about 437 pathways (72.8% with power transformation), with small numbers of remaining pathways allocated to a few clusters and a significant number of pathways (e.g. 22.7% with power transformation) considered as noise and ignored. This is due to the grouping mechanism of the algorithm, which tends to merge nearby pathways together iteratively as long as the pathways can be reached within a predefined distance from any pathway already included. As for the other clusters, they are more likely to be comprised of pathways at the extremes since those pathways are disconnected from the major part of the pathways and located in the tail of the distribution.

Table 2. Performance ranking of clustering algorithms for grouping the original and three transformed pathway ensembles based on five clustering validity indices (lower number means better performance).

Metric transformation	Clustering algorithm	Clustering validity index				
		SSE	DB	CH	DN	SL
None (original)	KM	1	1	1	2	1
	HC	2	1	2	1	2
	GMM	3	3	3	3	3
	SC	4	4	4	4	4
Power	KM	1	1	1	2	1
	HC	3	2	3	1	3
	GMM	2	3	2	3	2
	SC	4	4	4	4	4
Standardisation	KM	1	2	1	2	2
	HC	3	1	3	1	1
	GMM	2	4	2	3	3
	SC	4	3	4	4	4
Sectoral standardisation	KM	1	2	1	2	1
	HC	3	1	3	1	3
	GMM	2	3	2	3	2

* KM: K-means; HC: hierarchical clustering; GMM: Gaussian mixture model; SC: spectral clustering; SSE: sum of squared error index; DB: Davies-Bouldin index; CH: Calinski-Harabasz index; DN: Dunn; SL: Silhouette index

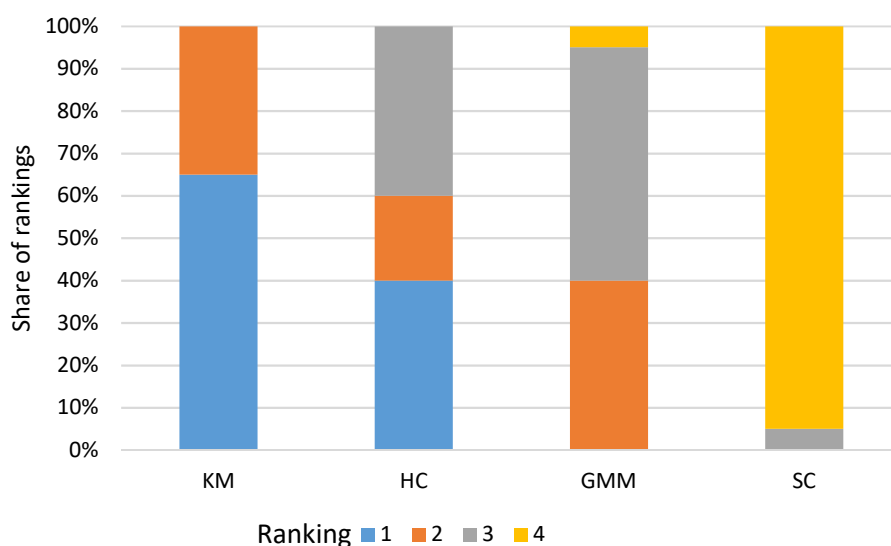


Figure 6. Share of rankings of clustering algorithms considering all four pathway ensembles based on four clustering validity indices (lower ranking number means better performance), showing KM is likely to outperform other algorithms in most cases.

Various clustering algorithms can lead to different pathway groupings, as illustrated in **Figure 7**, which presents the standard deviation of the number of pathways in the 5 identified clusters, for each combination of a clustering algorithm and a transformation metric. Lower standard deviation means that the distribution of the number of pathways across the different clusters is more uniform. Note that we focus here on five clusters of pathways, as the SSE analysis suggests that marginal benefits diminish with a larger number of clusters (see Appendix B).

KM and GMM have a similar distribution of numbers of pathways in clusters, with usually a much lower variation across clusters than with HC and SC. This is because KM tries to minimise the total distance between data points and their cluster centroids; as a result, and due to the pathways being fairly evenly distributed in the feature space, every cluster tends to cover a similar range of pathways so that the total distance is minimised. GMM, on the other hand, matches multiple Gaussian distributions to the distribution of pathways and

thus is less likely to group extreme pathways into an individual cluster. The numbers of grouped pathways are thus also more uniform.

Standardisation transformation amplifies the influence of the fluctuation of minor end-use metrics to make the distribution of pathways much more uneven. Groupings of HC and SC are thus affected by the extreme cases to form 4 and 3 clusters with less than 2 pathways respectively, whereas KM and GMM still have fairly even distributions of pathways across clusters. Due to the adopted merge strategy of HC, i.e. complete-linkage, only the maximum distance between boundary pathways (i.e. the two most dissimilar pathways in two clusters being merged) in two subsets of pathways is taken into account and HC therefore is more likely to leave out a smaller number of pathways in edge areas. Nonetheless, these groupings seem to be reasonable, as suggested by validity indices DB, DN, and SL, which also take extreme distances between pathways into account, in **Table 2** and Appendix B.3.

Even with the highly distorted distribution of standardised pathways, KM is still able to catch the extreme cases, with one cluster comprised of only 2 pathways, and to group pathways properly, as validity indices in **Table 2** show the performance of KM is either the best or the second best among all clustering algorithms.

As for SC, the significantly twisted distribution of standardised pathways might create higher dissimilarity between a few pathways and the major part of pathways; so that SC cuts those pathways into separate clusters. Nonetheless, the separations are still not ideal, as shown by the validity indices in **Table 2** and Appendix B.3.

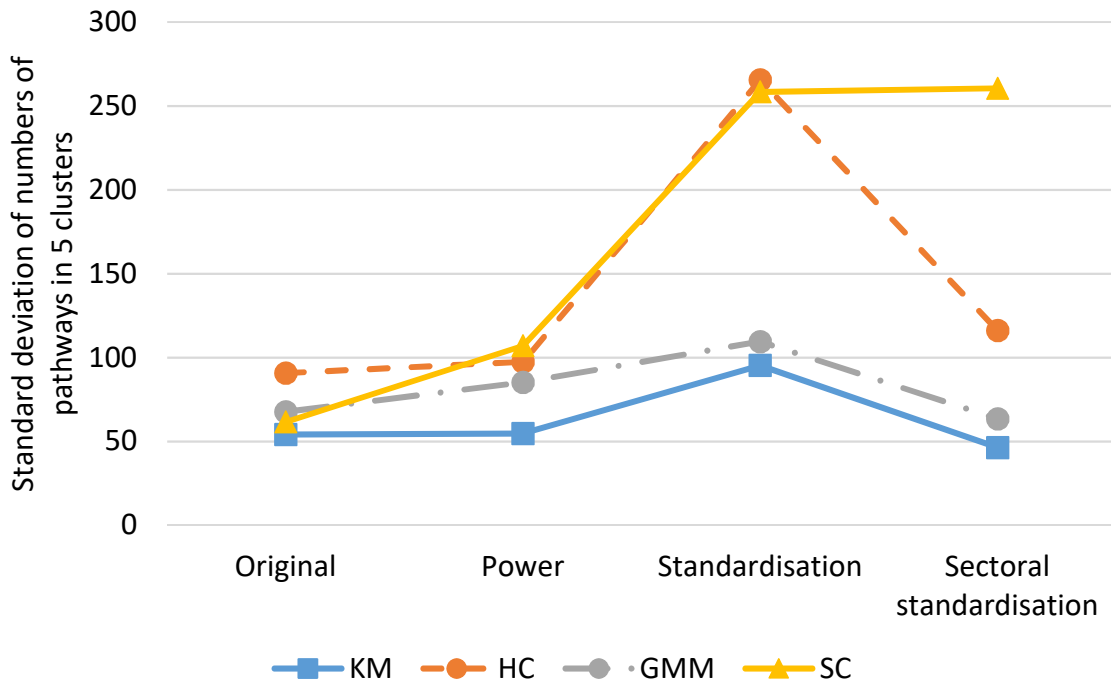


Figure 7 Standard deviation of numbers of pathways in 5 clusters identified by various algorithms for four considered metric transformation approaches to demonstrate the sensitivity of clustering algorithms to the distribution of proximity between transformed pathways. This result shows KM is the most robust algorithm.

4.4 Characteristics of representative pathways

KM, the most robust clustering algorithm for our specific dataset, is then adopted to characterise the pathway ensembles based on the sectoral-transformed metrics. This transformation was chosen as it can stress the importance of technologies in each sector in a balanced way. As shown in **Table 3**, the centroid pathways, i.e. representative pathways, of the five identified clusters represent the most distinctive combinations of technology metrics to decarbonise the whole energy system by 2050. The plus and minus signs represent the deviation level of mean metrics of pathways in *clusters* from the corresponding mean metrics of *all* pathways in 2030 and 2050. The variations in the industrial sector are ignored here as those are relative low in 2030 and negligible in 2050. The reader should note that the deviation of metrics in the table is based on the original scale rather than on the transformed scale that was used for the clustering.

Cluster 1 (“low carbon power generation”) suggests that more oil-based transport technologies can be used prior to 2050, if the building sector contributes more to mitigation

through increased use of bioenergy for heating and cooking in 2030. Less CCS-related technologies are deployed for electricity generation, which instead sees more wind and RE generation, and hydrogen production in 2050, possibly due to higher cost assumption of CCS under the pathways in this cluster.

Alternatively, as found in cluster 2 (“low carbon power generation with district heating and hydrogen cars”), district heating is deployed early on and to an even greater extent by 2050. This is augmented with higher solar thermal usage in 2050 to decarbonise the building sector. Furthermore, the higher production of hydrogen from coal and gas with CCS enables the replacement of oil cars with hydrogen vehicles.

Cluster 3 (“generation with CCS and wind, along with transport electrification”) illustrates yet another decarbonisation strategy to decarbonise the transport sector, with early and high level deployment of EVs. More electricity is generated with CCS-based power plants and wind turbines to meet the increased demand from EVs, while less hydrogen is required.

In cluster 4 (“more CCS for generation and extremely high BECCS for hydrogen production”), more bioenergy (only in 2030) and electricity-based end use technologies, such as heat pumps, are adopted to reduce the requirement of district heating in the building sector. Furthermore, much more hydrogen is produced through biomass gasification with CCS (BECCS) which contributes to negative GHG emissions. As a result of these negative emissions, more oil vehicles can still be used by 2050.

Finally, cluster 5 (“extremely high generation with CCS and high BECCS for hydrogen production”) is very similar to cluster 4, except that there is more gas consumption in 2030 and much more electricity is generated with CCS-fitted power plants.

Table 3. Deviation level of mean metrics of pathways in identified clusters by KM from the corresponding mean metrics of all pathways in 2030 and 2050, showing the characteristics of the identified representative pathways in clusters.

Sector		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
		2030	2050	2030	2050	2030	2050	2030	2050	2030	2050
Building	BIO	+				-		+		-	
	DH			+	++	+		-	--	-	--
	ELC				-				+		+
	GAS	-		-				-		+	
	SOL				+						

Electricity	CCS		---		---		++	-	++		+++	
	FOS	-				+						
	NUC		+++	+	+++		-		--	-	---	
	ORE	+	++		++				--		--	
	WND		++		++		++		--		---	
Hydrogen	BCCS		-		-		---	+	+++		++	
	CCCS				+		+	-	---		-	
	GCCS		-		+		+		-		-	
Transport	AV	GAS	--		+						+	
		OIL	++		-	-				+	-	+
	CAR	ELC				-	+	+++		--		-
		H2		+		++		-		-		-
		OIL	+		+		--	---	+	+++	+	+++
	HGV	GAS		-								
		OIL		+								
	LGV	ELC				+				-		-
		H2				+						
OIL		+			--	+			+	-	+	

Note: BIO: bioenergy; DH: district heating; ELC: electricity; GAS: gas; SOL: solar thermal; FOS: fossil fuel; NUC: nuclear; ORE: other renewable energy; WND: wind; H2: hydrogen; BCCS: bioenergy+CCS; CCCS: coal+CCS; GCCS: gas+CCS; AV: aviation; CAR: passenger car; HGV: heavy-good vehicle; LGV: light-good vehicle; +/-: positive/negative variation less than 5 TWh; ++/--: positive/negative variation about 10 TWh; +++/---: positive/negative variation about or more than 15 TWh.

The deviation of metrics in the five identified clusters also shows that some technologies have strong interdependencies, i.e. increasing/decreasing together or being substituted by each other. For instance, the increase of electricity generation with CCS could lead to the decrease of electricity generation with renewable energy and nuclear. The relationships between technology metrics across pathways are further described in **Figure 8** and **Figure 9**, which illustrate the differences between each pathway and the average pathway by metric in 2030 and 2050 respectively. Each point represents a relationship between two metrics of a pathway, with the colour indicating the cluster to which the pathway belongs. Stacked bar charts on the diagonal represent the distribution of metrics of different clusters. For example, the stacked bar chart circled in **Figure 8** shows that gas consumption in the building sector (BLD-GAS) shows two peaks in 2030, with clusters distributing almost evenly at these points.

Stronger correlations between metrics are found more frequently in 2050 than in 2030. In 2030, only the building sector has obvious trade-offs between technology metrics using different fuels for heating and cooking, i.e. gas usage can be replaced by bioenergy or district heating. Similar trade-offs in the building sector can also be observed in 2050. The

heating provision can be decarbonised either with more district heating or via electrification, using technologies such as heat pumps and electric storage heaters. Moreover, in 2050, increased electricity generation with CCS is likely to reduce the need of electricity from nuclear, wind and other renewable energy sources. Meanwhile, wind power and other renewable energy seem to have a positive correlation.

In 2050, hydrogen production by coal gasification with CCS (H2-CCCS) has a clear negative correlation with that by BECCS (H2-BCCS). Moreover, the way that hydrogen is produced is likely to influence the vehicle stock in 2050. Hydrogen production by BECCS can lead to negative GHG emissions, meaning more oil-based vehicles can remain on the road by 2050. Otherwise, oil-based vehicles are replaced by more EVs, as hydrogen is produced from coal with CCS.

Figure 8 and **Figure 9** also show the KM has successfully divided pathways into clear groups on some metrics, such as EVs (TCAR-ELC) and oil-based vehicles (TCAR-OIL and TLGV-OIL) in 2050. However, on most metrics, clear separations between clusters cannot be found due to the complicated interactions among metrics. This is also the reason why a proper clustering algorithm should be applied to identify distinct pathways as presented just before.

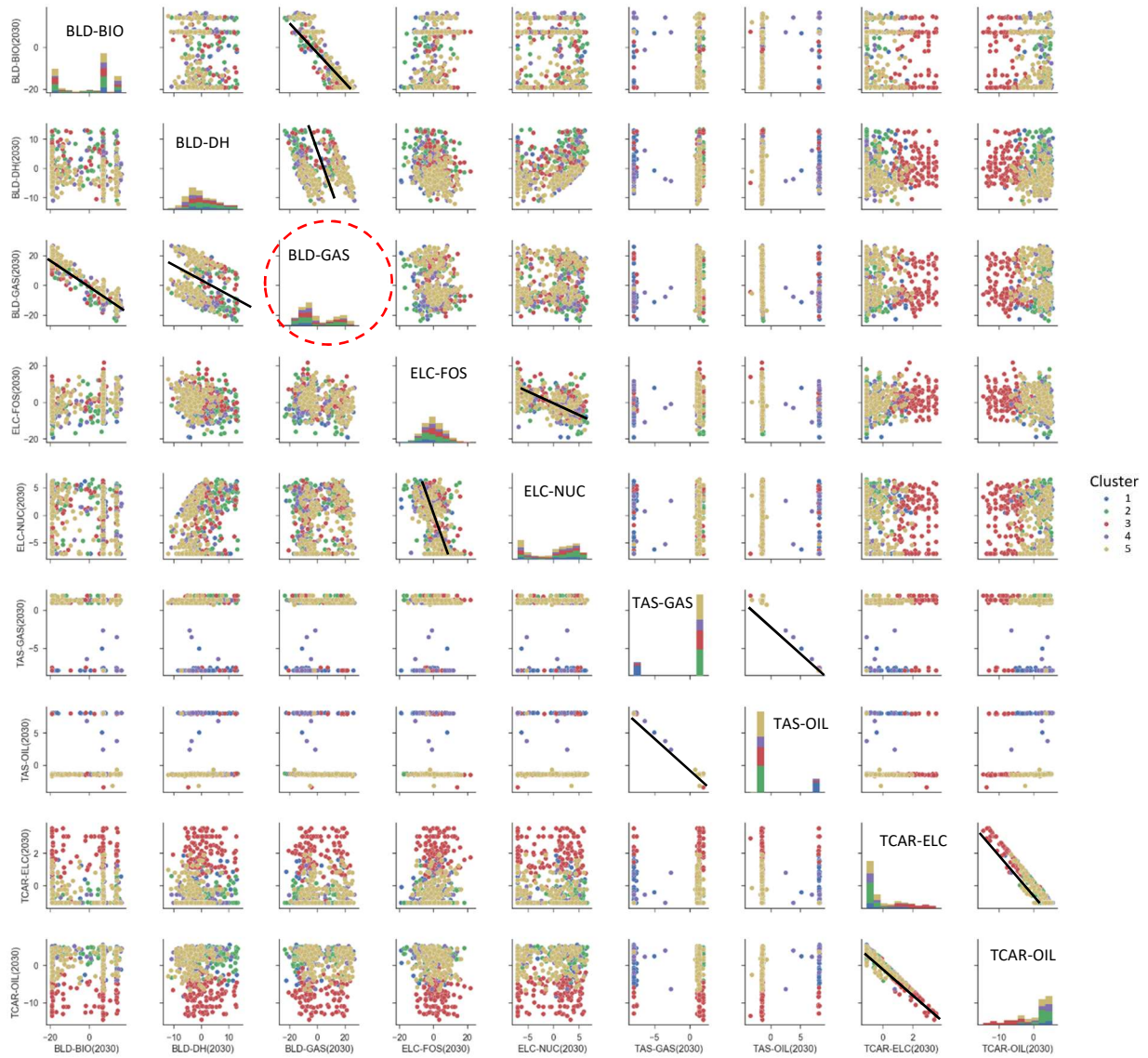


Figure 8. Scatter plots for technology metrics with higher variances in 2030 (lines indicate the obvious correlations between metrics) to show the potential substitutions between technologies across pathways.

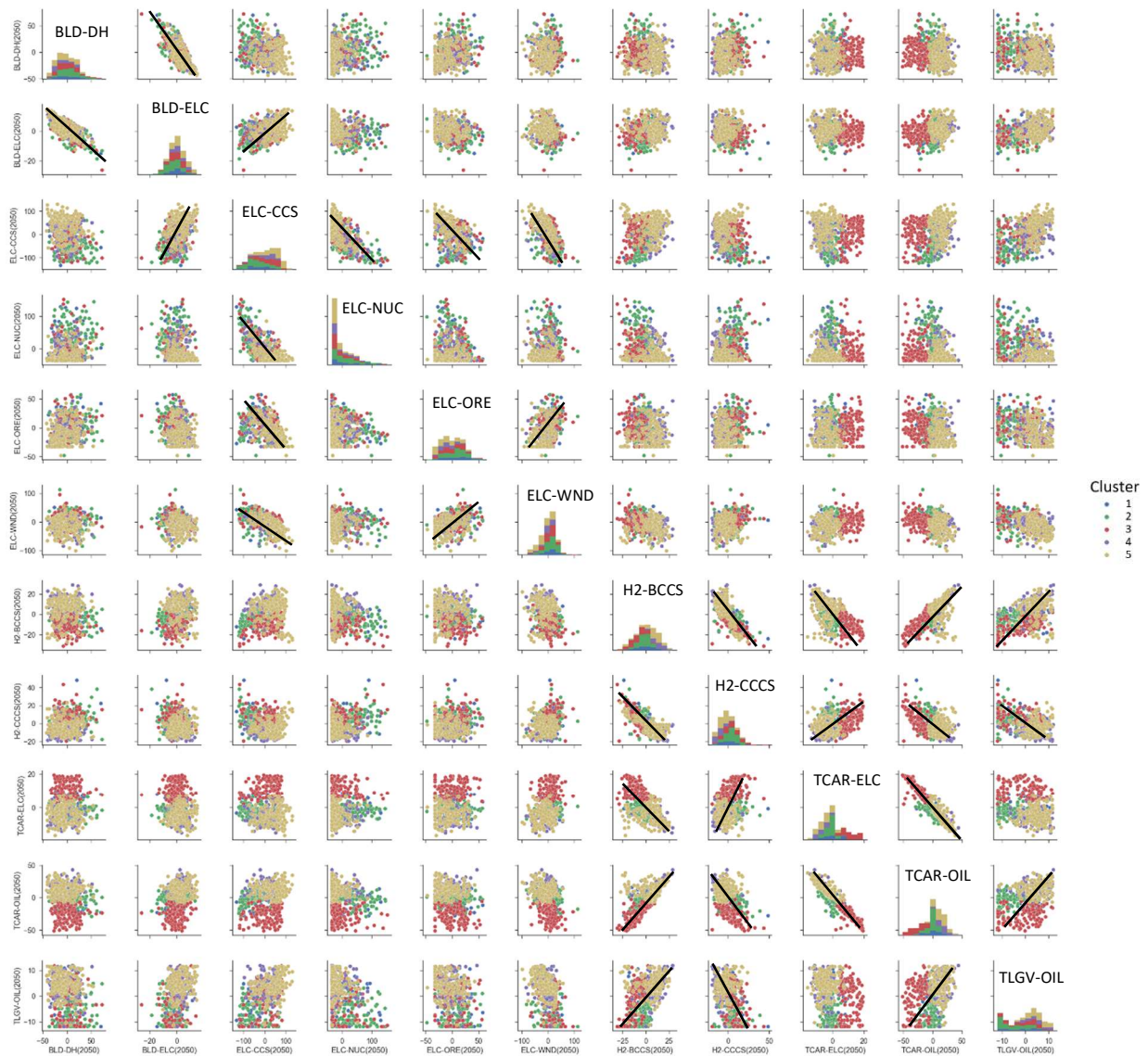


Figure 9. Scatter plots for technology metrics with higher variances in 2050 (lines indicate the obvious correlations between metrics) to show the potential substitutions between technologies across pathways.

4.5 Different clustering results of the other clustering algorithms

To understand how the choice of clustering algorithm can affect the identification of representative pathways, HC, GMM, and SP were also applied to group the sectoral standardised pathways into five clusters. The representative pathways identified by those algorithms are illustrated in **Table D.1** to **Table D.3** in Appendix D. In general, similar trade-offs between metrics within individual sectors can be clearly observed from the pathways identified by various algorithms. For instance, the trade-off between power generation with CCS and low-carbon power generation in the electricity sector can be found in the

representative pathways identified by almost all the considered algorithms, except SC doesn't suggest centroid pathways with a similar trade-off, as shown in **Table D.3**. Nonetheless, not all scenarios revealed by the representative pathways from KM analysis can be found in those for the other considered algorithms. For example, the scenario with high generation with CCS and strong transport electrification (i.e. cluster 3 from KM) is not identified by the other algorithms. Furthermore, low-carbon power generation scenario (i.e. cluster 1 from KM) is only identified by KM, not the other algorithms.

Therefore, the application of various clustering algorithms can lead to very different representative pathways to characterise pathway ensembles. It is thus crucial to adopt a suitable algorithm for this task based on the distribution characteristics of pathways.

5 Conclusions

Clustering algorithms provide a powerful approach to grouping data points with high dimensional features into clusters based on the similarity between those data points. This study thus adopted clustering algorithms to group 600 decarbonisation pathways (based on 2030 and 2050) reflecting multiple uncertainties, to identify distinct characteristics of representative pathways. Four transformation approaches, including original scale, power-transformation, standardisation and sectoral standardisation, were applied to pre-process the pathways to strengthen the variances of individual metrics. Transformed pathways were then grouped with five clustering algorithms and the results were compared using five cluster validity indices. K-means was found to be the best algorithm for our pathway dataset as it consistently outperformed the other considered algorithms. This finding was generally robust across the transformation approaches and validity indexes used. Furthermore, sectoral standardisation, a novel transformation approach proposed in this study, was shown as a useful transformation approach for revealing the distinct variances of metrics in the end-use sectors for clustering, in addition to mitigating the influence of a few dominant metrics in the power sector.

The application of clustering algorithms to characterise long-term pathways has been shown to find distinct, yet representative pathways, from a large set of pathways. This allows decision makers to make robust decisions, when exposed to the wide range of

uncertainties reflected by the original scenario ensemble. Furthermore, the approach is flexible in grouping pathways with different pathway transformations, e.g. sectoral standardisation, to emphasise trade-offs between metrics in pathways from various perspectives.

However, the applied clustering algorithm should be carefully chosen since various clustering algorithms can, based on our results, lead to very different representative pathways. Even though k-means could be the preferred choice for pathway clustering when the distribution of pathways is continuous and condensed, the results based on the standardisation transformation suggest that hierarchical clustering algorithm may perform better, if the distribution of pathways is less regular and sparser.

Pathway distribution is, thus, a key factor influencing the performance of clustering algorithms, which means that our findings are linked to the specific model based dataset we have used here. For future studies, broader, and differently shaped parameter distributions could also be imposed in the pathway generation process, to see how this affects the performance of different clustering algorithms. Pathway ensembles generated by different models can also be put together for further testing. In addition, other types of clustering algorithms, such as kernel k-means and support vector machine, can be applied to characterise pathways to further validate the robustness of the findings.

Clustering algorithms can also be applied to group individual technology metrics directly, rather than pathways, to identify highly correlated metric sets, as done in [26] using hierarchical clustering. Finally, the proposed approach can easily be applied to analyse pathways generated by any national or global scale energy system models, not only the energy system model adopted in this study.

Acknowledgements

This research has received funding through REEM project from the European Union's Horizon 2020 research and innovation programme under grant agreement 691739.

References

- [1] United Nation Framework Convention on Climate Change. The Paris Agreement 2015.

- <https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement>.
- [2] Loulou R, Goldstein G, Kanudia A, Lettila A, Remme U. Documentation for the TIMES Model - Part I. Energy Technology Systems Analysis Programme; 2016.
- [3] Deane P, Chiodi A, Gallachóir BÓ. Report No . 199 Irish TIMES Energy Systems Model Phase 2. Environmental Protection Agency, Ireland; 2016.
- [4] Riekkola AK. National energy system modelling for supporting energy and climate policy decision making: The case of Sweden. Chalmers University of Technology, 2015.
- [5] Hughes N, Strachan N, Gross R. The structure of uncertainty in future low carbon pathways. *Energy Policy* 2013;52:45–54.
<https://doi.org/10.1016/j.enpol.2012.04.028>.
- [6] Nilsson M, Nykvist B. Governing the electric vehicle transition – Near term interventions to support a green energy economy. *Appl Energy* 2016;179:1360–71.
<https://doi.org/10.1016/j.apenergy.2016.03.056>.
- [7] van Bree B, Verbong GPJ, Kramer GJ. A multi-level perspective on the introduction of hydrogen and battery-electric vehicles. *Technol Forecast Soc Change* 2010;77:529–40. <https://doi.org/10.1016/j.techfore.2009.12.005>.
- [8] Knez M, Zevnik GK, Obrecht M. A review of available chargers for electric vehicles: United States of America, European Union, and Asia. *Renew Sustain Energy Rev* 2019;109:284–93. <https://doi.org/10.1016/j.rser.2019.04.013>.
- [9] Zubi G, Dufo-López R, Carvalho M, Pasaoglu G. The lithium-ion battery: State of the art and future perspectives. *Renew Sustain Energy Rev* 2018;89:292–308.
<https://doi.org/10.1016/j.rser.2018.03.002>.
- [10] Deuten S, Gómez Vilchez JJ, Thiel C. Analysis and testing of electric car incentive scenarios in the Netherlands and Norway. *Technol Forecast Soc Change* 2020;151:119847. <https://doi.org/10.1016/j.techfore.2019.119847>.
- [11] IEA. Global EV Outlook 2019 – Scaling up the transition to electric mobility 2019. <https://www.iea.org/reports/global-ev-outlook-2019> (accessed February 24, 2020).
- [12] Strielkowski W, Štreimikienė D, Bilan Y. Network charging and residential tariffs: A

- case of household photovoltaics in the United Kingdom. *Renew Sustain Energy Rev* 2017;77:461–73. <https://doi.org/10.1016/j.rser.2017.04.029>.
- [13] BEIS. Solar photovoltaics deployment - GOV.UK 2020. <https://www.gov.uk/government/statistics/solar-photovoltaics-deployment> (accessed February 24, 2020).
- [14] Mei H, Li YP, Suo C, Ma Y, Lv J. Analyzing the impact of climate change on energy-economy-carbon nexus system in China. *Appl Energy* 2020;262:114568. <https://doi.org/10.1016/j.apenergy.2020.114568>.
- [15] Park CK, Kim HJ, Kim YS. A study of factors enhancing smart grid consumer engagement. *Energy Policy* 2014;72:211–8. <https://doi.org/10.1016/j.enpol.2014.03.017>.
- [16] Tattini J, Gargiulo M, Karlsson K. Reaching carbon neutral transport sector in Denmark – Evidence from the incorporation of modal shift into the TIMES energy system modeling framework. *Energy Policy* 2018;113:571–83. <https://doi.org/10.1016/j.enpol.2017.11.013>.
- [17] Shi J, Chen W, Yin X. Modelling building’s decarbonization with application of China TIMES model. *Appl Energy* 2016;162:1303–12. <https://doi.org/10.1016/j.apenergy.2015.06.056>.
- [18] Silva Herran D, Tachiiri K, Matsumoto K. Global energy system transformations in mitigation scenarios considering climate uncertainties. *Appl Energy* 2019;243:119–31. <https://doi.org/10.1016/j.apenergy.2019.03.069>.
- [19] Panos E, Kober T, Wokaun A. Long term evaluation of electric storage technologies vs alternative flexibility options for the Swiss energy system. *Appl Energy* 2019;252:113470. <https://doi.org/10.1016/j.apenergy.2019.113470>.
- [20] Fais B, Keppo I, Zeyringer M, Usher W, Daly H. Impact of technology uncertainty on future low-carbon pathways in the UK. *Energy Strateg Rev* 2016;13–14:154–68. <https://doi.org/10.1016/j.esr.2016.09.005>.
- [21] DeCarolus JF, Babae S, Li B, Kanungo S. Modelling to generate alternatives with an energy system optimization model. *Environ Model Softw* 2016;79:300–10.

- <https://doi.org/10.1016/j.envsoft.2015.11.019>.
- [22] Price J, Keppo I. Modelling to generate alternatives: A technique to explore uncertainty in energy-environment-economy models. *Appl Energy* 2017;195:356–69. <https://doi.org/10.1016/j.apenergy.2017.03.065>.
- [23] Li FGN, Trutnevyte E. Investment appraisal of cost-optimal and near-optimal pathways for the UK electricity sector transition to 2050. *Appl Energy* 2017;189:89–109. <https://doi.org/10.1016/j.apenergy.2016.12.047>.
- [24] Yue X, Pye S, DeCarolis J, Li FGN, Rogan F, Gallachóir B. A review of approaches to uncertainty assessment in energy system optimization models. *Energy Strateg Rev* 2018;21:204–17. <https://doi.org/10.1016/j.esr.2018.06.003>.
- [25] Pye S, Sabio N, Strachan N. An integrated systematic analysis of uncertainties in UK energy transition pathways. *Energy Policy* 2015;87:673–84. <https://doi.org/10.1016/j.enpol.2014.12.031>.
- [26] Pye S, Li PH, Keppo I, O’Gallachoir B. Technology interdependency in the United Kingdom’s low carbon energy transition. *Energy Strateg Rev* 2019;24:314–30. <https://doi.org/10.1016/j.esr.2019.04.002>.
- [27] Fragkos P, Kouvaritakis N, Capros P. Incorporating Uncertainty into World Energy Modelling: the PROMETHEUS Model. *Environ Model Assess* 2015;20:549–69. <https://doi.org/10.1007/s10666-015-9442-x>.
- [28] Pizarro-Alonso A, Ravn H, Münster M. Uncertainties towards a fossil-free system with high integration of wind energy in long-term planning. *Appl Energy* 2019;253:113528. <https://doi.org/10.1016/j.apenergy.2019.113528>.
- [29] Moksnes N, Rozenberg J, Broad O, Taliotis C, Howells M, Rogner H. Determinants of energy futures—a scenario discovery method applied to cost and carbon emission futures for South American electricity infrastructure. *Environ Res Commun* 2019;1:025001. <https://doi.org/10.1088/2515-7620/AB06DE>.
- [30] Tan P-N, Steinbach M, Karpatne A, Kumar V. Introduction to data mining. Second edi. Pearson; 2017.
- [31] Jain AK. Data clustering: 50 years beyond K-means. *Pattern Recognit Lett*

- 2010;31:651–66. <https://doi.org/10.1016/j.patrec.2009.09.011>.
- [32] Ma Z, Yan R, Li K, Nord N. Building energy performance assessment using volatility change based symbolic transformation and hierarchical clustering. *Energy Build* 2018;166:284–95. <https://doi.org/10.1016/j.enbuild.2018.02.015>.
- [33] Schütz T, Schraven MH, Fuchs M, Remmen P, Müller D. Comparison of clustering algorithms for the selection of typical demand days for energy system synthesis. *Renew Energy* 2018;129:570–82. <https://doi.org/10.1016/j.renene.2018.06.028>.
- [34] Tang F, Kusiak A, Wei X. Modeling and short-term prediction of HVAC system with a clustering algorithm. *Energy Build* 2014;82:310–21. <https://doi.org/10.1016/j.enbuild.2014.07.037>.
- [35] Liao N, He Y. Exploring the effects of influencing factors on energy efficiency in industrial sector using cluster analysis and panel regression model. *Energy* 2018;158:782–95. <https://doi.org/10.1016/j.energy.2018.06.049>.
- [36] Unternährer J, Moret S, Joost S, Maréchal F. Spatial clustering for district heating integration in urban energy systems: Application to geothermal energy. *Appl Energy* 2017;190:749–63. <https://doi.org/10.1016/j.apenergy.2016.12.136>.
- [37] Chidean MI, Caamaño AJ, Ramiro-Bargueño J, Casanova-Mateo C, Salcedo-Sanz S. Spatio-temporal analysis of wind resource in the Iberian Peninsula with data-coupled clustering. *Renew Sustain Energy Rev* 2018;81:2684–94. <https://doi.org/10.1016/j.rser.2017.06.075>.
- [38] Papadopoulos S, Bonczak B, Kontokosta CE. Pattern recognition in building energy performance over time using energy benchmarking data. *Appl Energy* 2018;221:576–86. <https://doi.org/10.1016/j.apenergy.2018.03.079>.
- [39] Yilmaz S, Chambers J, Patel MK. Comparison of clustering approaches for domestic electricity load profile characterisation - Implications for demand side management. *Energy* 2019;180:665–77. <https://doi.org/10.1016/j.energy.2019.05.124>.
- [40] Boudet HS, Flora JA, Armel KC. Clustering household energy-saving behaviours by behavioural attribute. *Energy Policy* 2016;92:444–54. <https://doi.org/10.1016/j.enpol.2016.02.033>.

- [41] Arias MB, Bae S. Electric vehicle charging demand forecasting model based on big data technologies. *Appl Energy* 2016;183:327–39.
<https://doi.org/10.1016/j.apenergy.2016.08.080>.
- [42] Dong L, Wang L, Farhan S, Gao S, Liao X. Wind power day-ahead prediction with cluster analysis of NWP. *Renew Sustain Energy Rev* 2016;60:1206–12.
<https://doi.org/10.1016/j.rser.2016.01.106>.
- [43] Sun G, Jiang C, Cheng P, Liu Y, Wang X, Fu Y, He Y. Short-term wind power forecasts by a synthetical similar time series data mining method. *Renew Energy* 2018;115:575–84. <https://doi.org/10.1016/j.renene.2017.08.071>.
- [44] Csereklyei Z, Thurner PW, Langer J, Küchenhoff H. Energy paths in the European Union: A model-based clustering approach. *Energy Econ* 2017;65:442–57.
<https://doi.org/10.1016/j.eneco.2017.05.014>.
- [45] Saxena A, Prasad M, Gupta A, Bharill N, Patel OP, Tiwari A, Er MJ, Ding W, Lin C-T. A review of clustering techniques and developments. *Neurocomputing* 2017;267:664–81. <https://doi.org/10.1016/j.neucom.2017.06.053>.
- [46] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: Machine learning in Python. *J Mach Learn Res* 2011;12:2825–30.
- [47] Heaton C. Modelling Low-Carbon Energy System Designs with the ETI ESME Model. 2014.
- [48] developers S. Scikit-learn User Guide-2.3 Clustering 2019.
- [49] Murphy KP. *Machine learning : a probabilistic perspective*. MIT Press; 2012.
- [50] Ghiassi N, Mahdavi A. Reductive bottom-up urban energy computing supported by multivariate cluster analysis. *Energy Build* 2017;144:372–86.
<https://doi.org/10.1016/j.enbuild.2017.03.004>.
- [51] Li R, Wang Z, Gu C, Li F, Wu H. A novel time-of-use tariff design based on Gaussian Mixture Model. *Appl Energy* 2016;162:1530–6.
<https://doi.org/10.1016/j.apenergy.2015.02.063>.

- [52] Bishop CM. Pattern recognition and machine learning. Springer; 2006.
- [53] Jianbo Shi, Malik J. Normalized cuts and image segmentation. *IEEE Trans Pattern Anal Mach Intell* 2000;22:888–905. <https://doi.org/10.1109/34.868688>.
- [54] Liu D, Wang J, Wang H. Short-term wind speed forecasting based on spectral clustering and optimised echo state networks. *Renew Energy* 2015;78:599–608. <https://doi.org/10.1016/j.renene.2015.01.022>.
- [55] Nadler B, Galun M. Fundamental limitations of spectral clustering. *Adv. Neural Inf. Process. Syst.*, 2007, p. 1017–24.
- [56] Arbelaitz O, Gurrutxaga I, Muguerza J. An extensive comparative study of cluster validity indices 2013;46:243–56. <https://doi.org/10.1016/j.patcog.2012.07.021>.
- [57] Cayla J, Maïzi N. Integrating household behavior and heterogeneity into the TIMES-Households model. *Appl Energy* 2015;139:56–67.

Appendix A. Definitions of technology metrics

Metric	Definition	Units
ELC-WND	Wind generation level	TWh
ELC-NUC	Nuclear generation level	TWh
ELC-CCS	CCS generation level	TWh
ELC-ORE	Other renewable generation level	TWh
ELC-FOS	Fossil generation level	TWh
BLD-BIO	Building bioenergy consumption	TWh
BLD-ELC	Building electricity consumption	TWh
BLD-GAS	Building gas consumption	TWh
BLD-OIL	Building oil consumption	TWh
BLD-DH	Building district heating consumption	TWh
BLD-SOL	Building solar energy consumption	TWh
H2-BCCS	H2 production by biomass gasification with CCS	TWh
H2-CCCS	H2 production by coal gasification with CCS	TWh
H2-ELC	H2 production by electrolysis	TWh
H2-GCCS	H2 production by gas (SMR) with CCS	TWh
H2-GAS	H2 production by gas (SMR)	TWh
IND-BIO	Industry bioenergy consumption	TWh
IND-COA	Industry coal consumption	TWh
IND-ELC	Industry electricity consumption	TWh
IND-GAS	Industry gas consumption	TWh
IND-H2	Industry hydrogen consumption	TWh
IND-OIL	Industry oil consumption	TWh
TAS-GAS	Aviation & shipping - gas	TWh
TAS-OIL	Aviation & shipping - oil	TWh
TAS-BFL	Aviation & shipping - biofuel	TWh
TCAR-ELC	Cars - electricity	TWh
TCAR-GAS	Cars - gas	TWh
TCAR-H2	Cars - H2	TWh
TCAR-OIL	Cars - oil	TWh
TCAR-BFL	Cars - biofuels	TWh
THGV-ELC	Heavy goods vehicles - electricity	TWh
THGV-GAS	Heavy goods vehicles - gas	TWh
THGV-H2	Heavy goods vehicles - H2	TWh
THGV-OIL	Heavy goods vehicles – oil	TWh
THGV-BFL	Heavy goods vehicles - biofuels	TWh
TLGV-ELC	Light goods vehicles - electricity	TWh
TLGV-H2	Light goods vehicles - H2	TWh
TLGV-OIL	Light goods vehicles - oil	TWh
TLGV-BFL	Light goods vehicles - biofuels	TWh

Appendix B. Comparison of performance of clustering algorithms using technology metrics with various transformations

We present below the validity indices as a function of the number of clusters for the four clustering algorithms and the four different transformations. The arrows in the vertical titles indicate the improvement direction of an index, e.g. upward arrow means the index is better if its value is larger.

B.1 No transformation

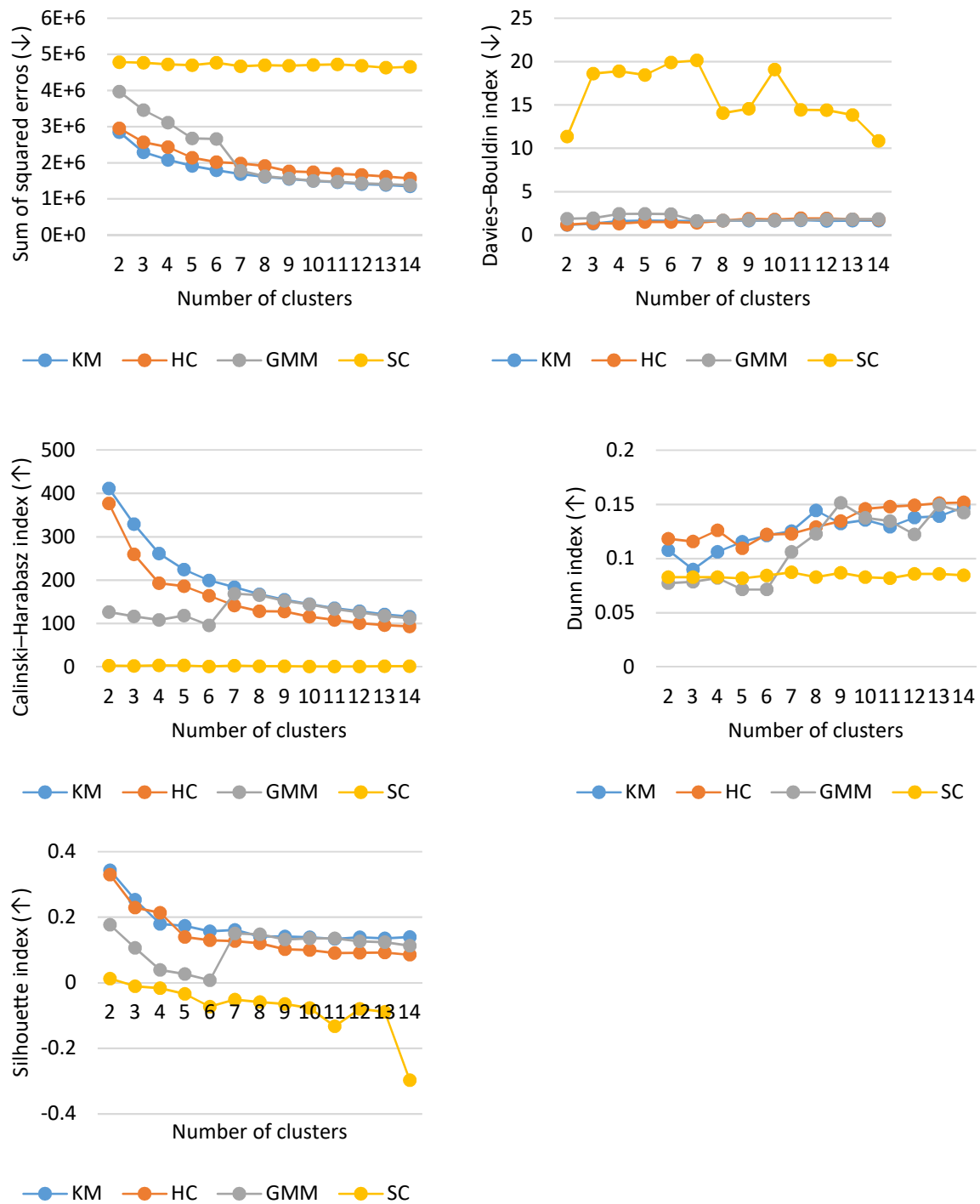


Figure B.1. Validity indices against various numbers of clusters of pathways for four clustering algorithms using technology metrics without transformation

B.2 Log-transformation

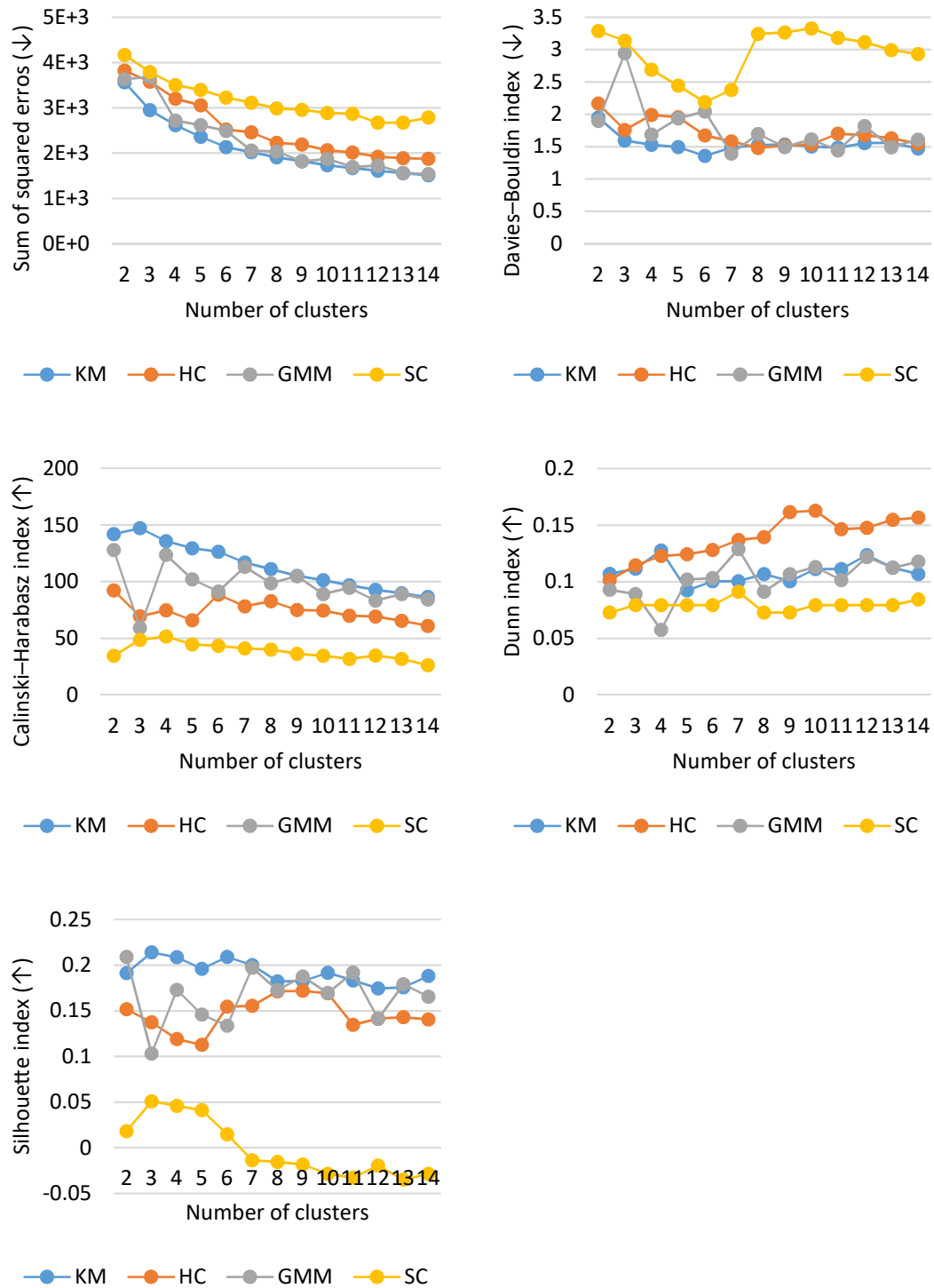


Figure B.2. Validity indices against various numbers of clusters of pathways for four clustering algorithms

B.3 Standardisation transformation

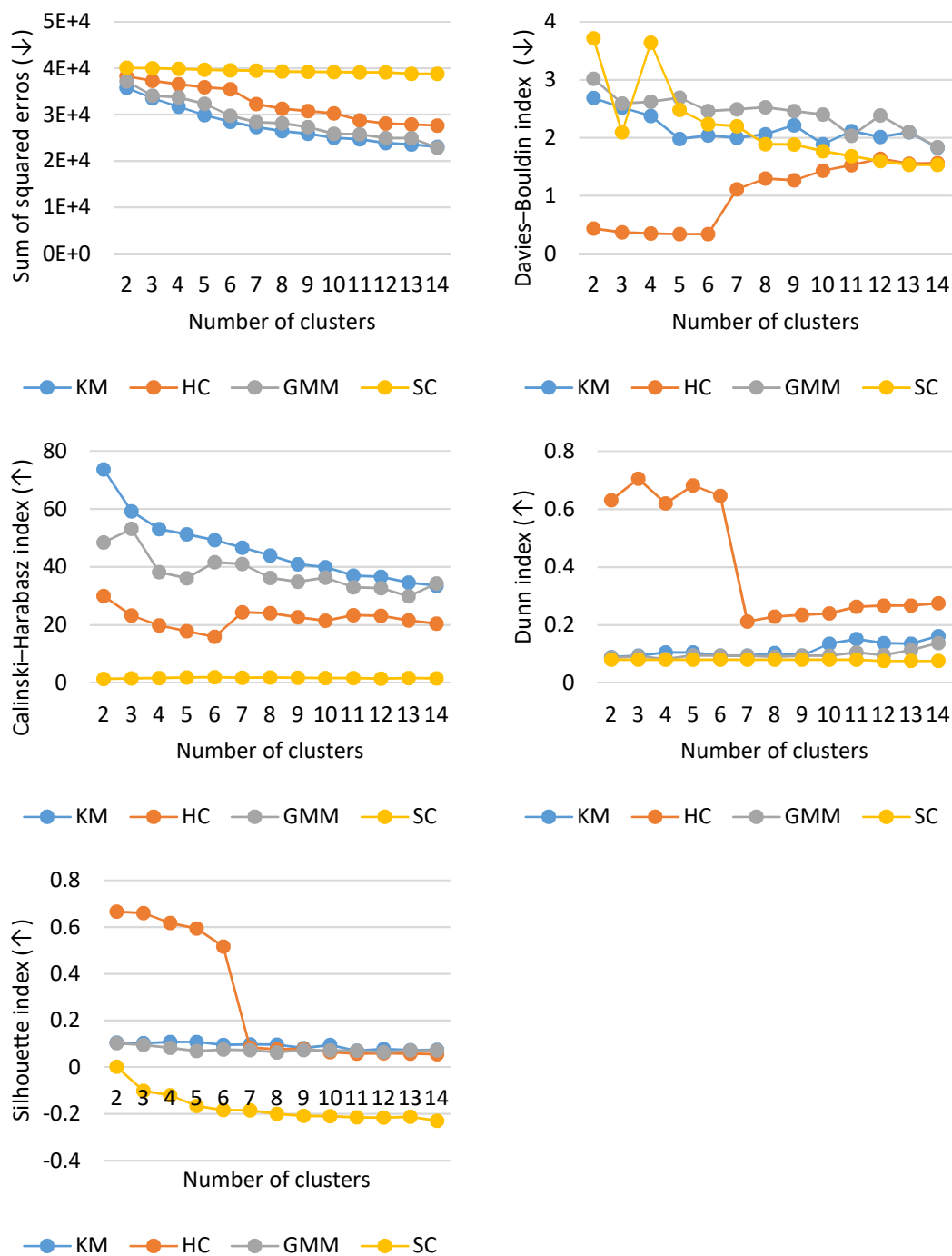


Figure B.3. Validity indices against various numbers of clusters of pathways for four clustering algorithms

B.4 Sectoral standardisation transformation

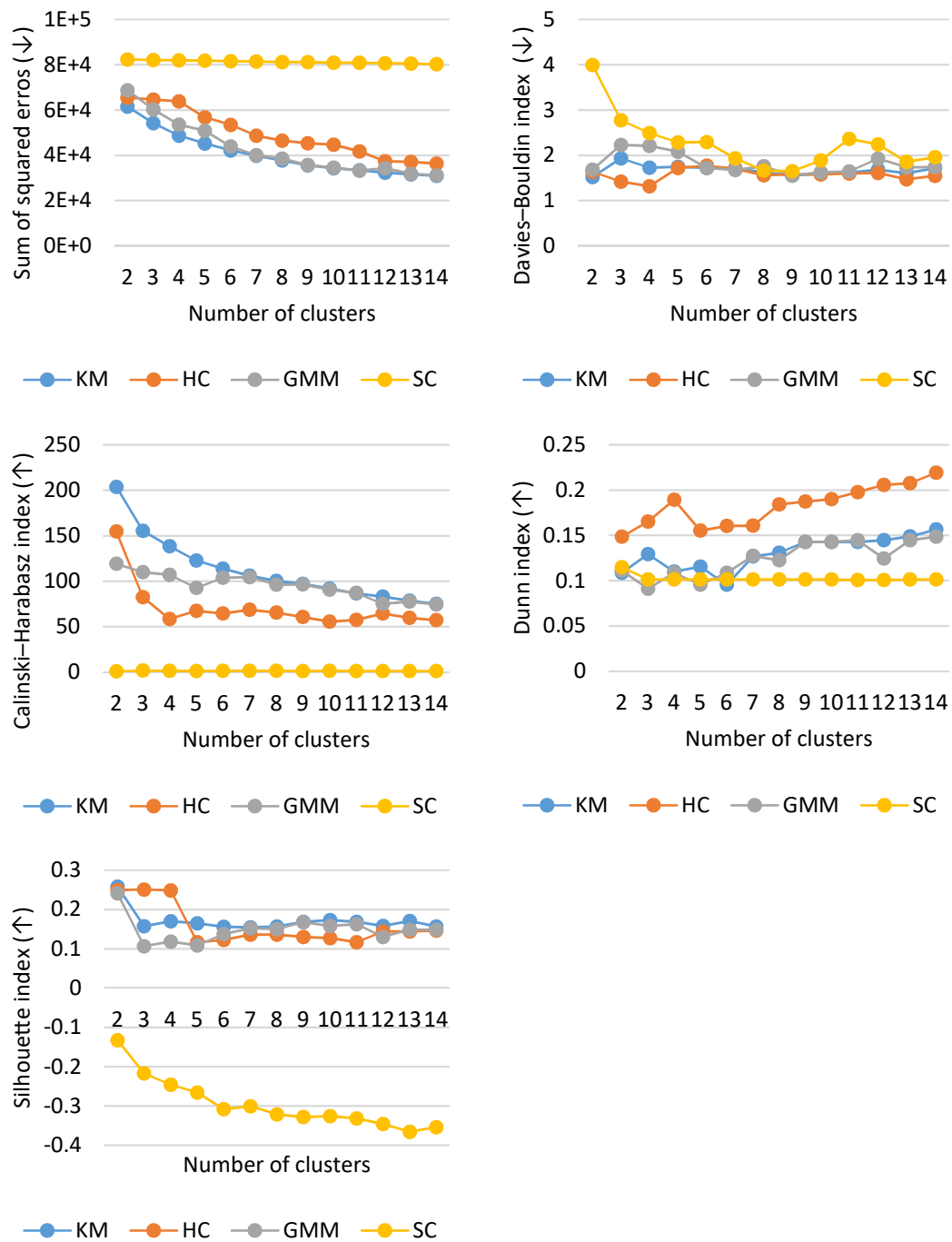


Figure B.4. Validity indices against various numbers of clusters of pathways for four clustering algorithms

Appendix C. Influence of metric transformations on k-means clustering analysis

The influence of metric transformations on the clustering of pathways is discussed below. For simplicity' sake, we show only clustering done with k-means (KM); those determined by other clustering algorithms would be broadly similar and KM was found in this study to be the most robust algorithm for clustering the pathways in our dataset.

The differences between the mean technology metrics of the pathways in the 5 clusters and the average metrics of all pathways in 2030 and 2050 are illustrated in **Figure C.1** to **Figure C.4**. It should be noted that technology metrics in those figures are on the original scales.

The major effects of the transformations on the grouped clusters are rescaling the boundaries of extreme metrics and shifting the signal strengths of metrics. For example, the extreme values of original metrics are in the range of 100 to -100 TWh, as shown in **Figure C.1**, whereas those values are much lower (between 60 and -60 TWh) after applying power-transformation, as shown in **Figure C.2**.

As the fluctuations of original metrics vary widely, the clustering results are thus dominated by the metrics with high variance, such as those in the electricity sector. The deviation from the average metrics of pathway groups in the end-use sectors are almost negligible, as shown in **Figure C.1**, since the proximity between pathways is most likely to be determined by a few metrics with higher variance.

Clustering results for the power-transformed pathways, on the other hand, strengthen the variance of some extreme cases in the end-use sectors by mitigating the influences of a few metrics with high variance. Variance of bioenergy consumption in the buildings in 2030 (BLD-BIO) and hydrogen production by gas (SMR) and CCS in 2050 (H2-GCCS) are thus more obvious in **Figure 3(b)** and **Figure C.2**.

In addition, clustering results for the standardised pathways emphasise the variance of every metric almost equally and thus leads to more obvious fluctuations of metrics in 2030 and 2050 than those for original pathways. For example, the variation of gas consumption of heavy good vehicles in 2050 (THGV-GAS) is almost zero, as shown in **Figure C.1**, while a higher variance can be found in one identified pathway cluster, as shown in **Figure C.3**.

Finally, sectoral standardisation strengthens the influence of metrics variance in individual sectors, as shown in **Figure 3(d)**, on the determination of proximity between pathways. Relatively high variance of metrics in a sector are thus more likely to be revealed, such as gas consumption of aviation in 2030 (TAS-GAS) and electricity consumption of cars in 2050 (TCAR-ELC), as shown in **Figure C.4**. Those variances are less obvious in the clustering results of original pathways (**Figure C.1**) but occasionally can be emphasised in the standardised case (**Figure C.3**). Moreover, the trade-off between technology metrics is less clear in the case with power-transformed pathways (**Figure C.2**), such as the replacement of oil cars (TCAR-OIL) with EVs (TCAR-ELC) in 2050 among clusters is missing.

Overall, as discussed above, grouping pathways with KM using sectoral-transformed pathways could be a better approach to identify distinct pathway groups in terms of the variations of metrics across pathways.

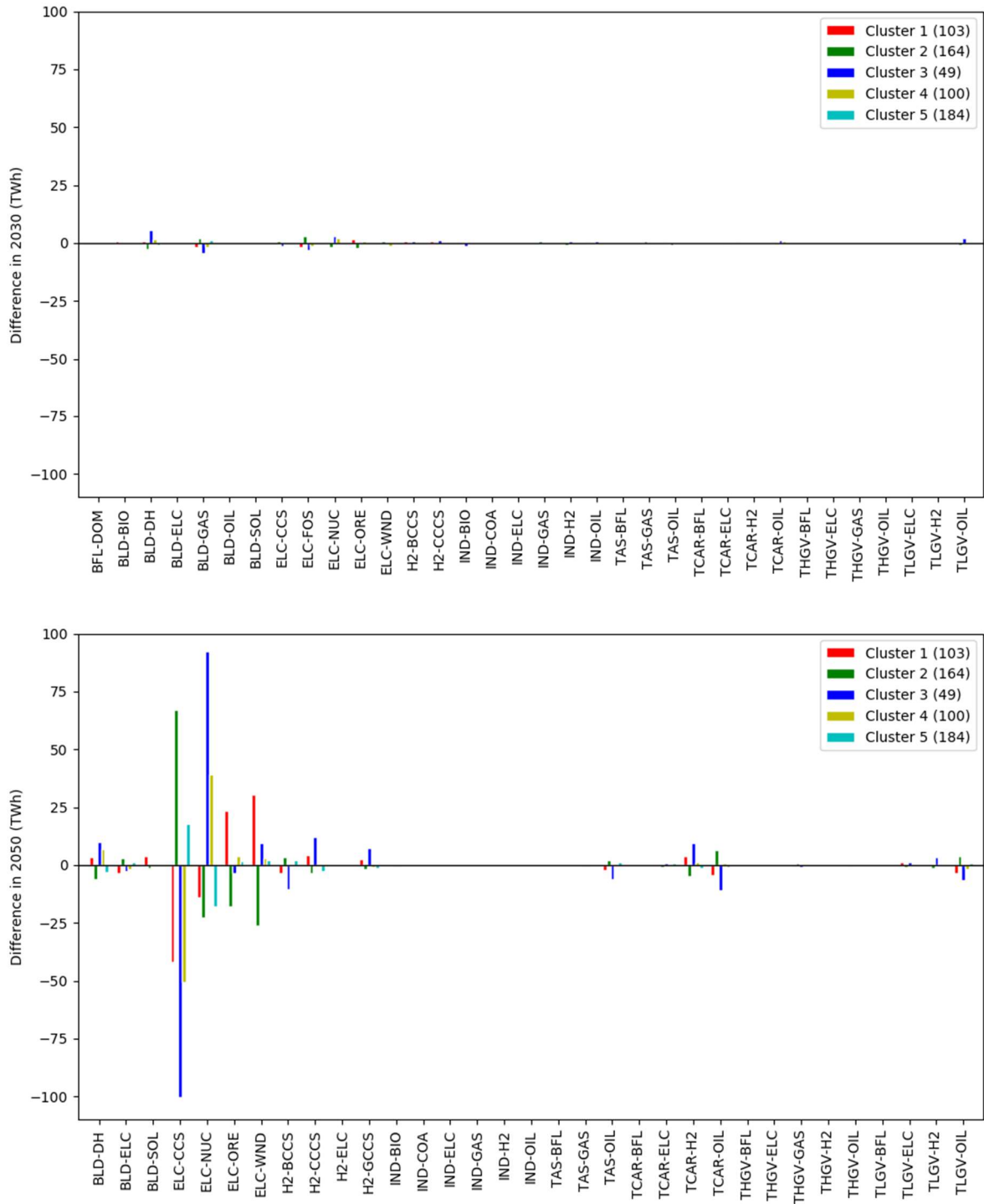


Figure C.1 Differences between the average metrics of all pathways and the mean metrics of clusters in 2030 and 2050

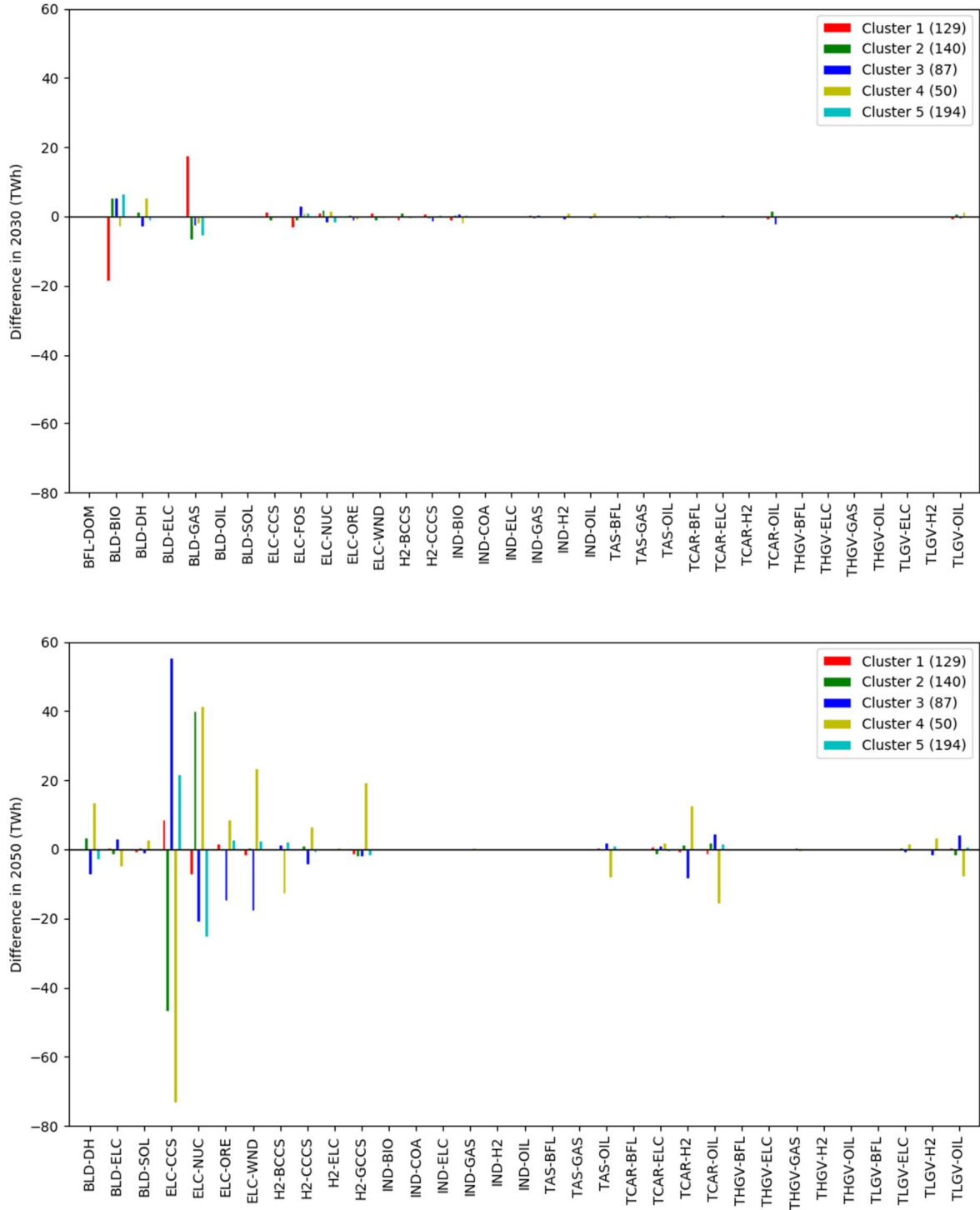


Figure C.2 Differences between the average metrics of all pathways and the mean metrics of clusters in 2030 and 2050 identified with power-transformed metrics

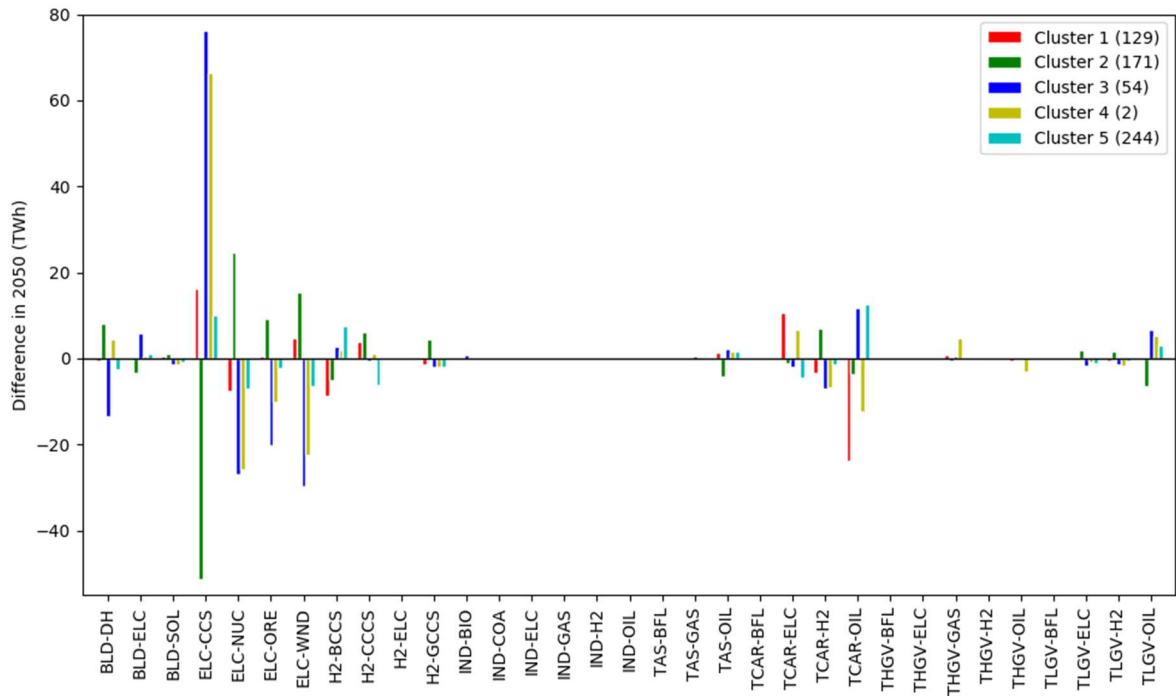
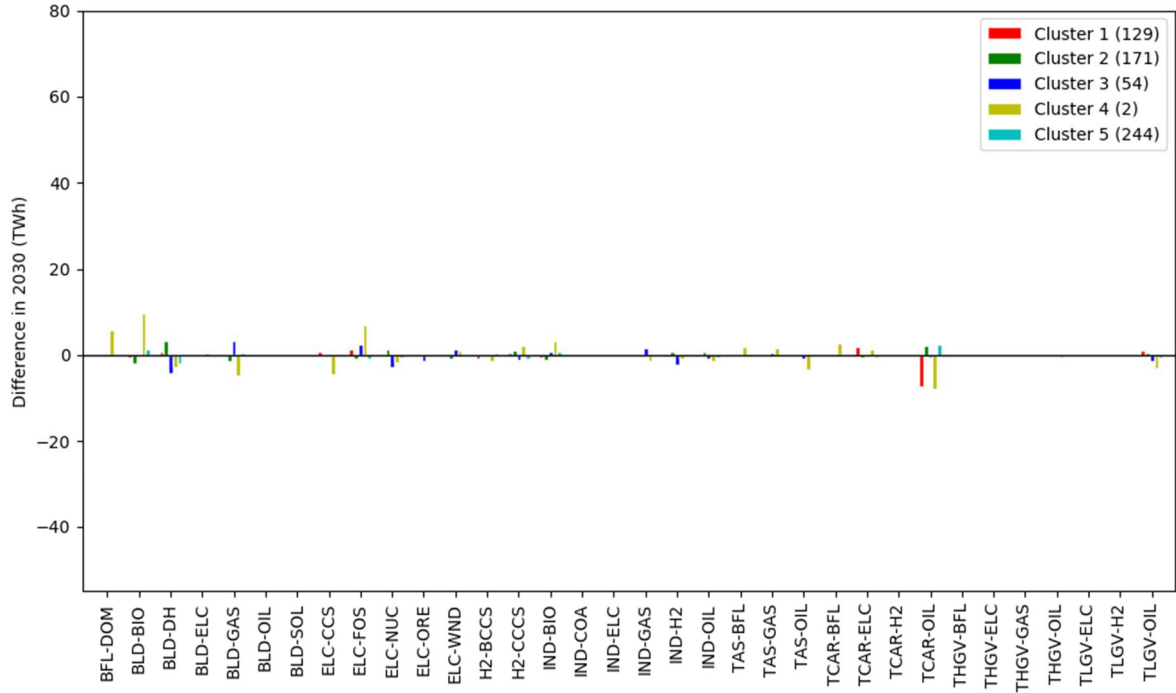


Figure C.3 Differences between the average metrics of all pathways and the mean metrics of clusters in 2030 and 2050 identified with standardised metrics

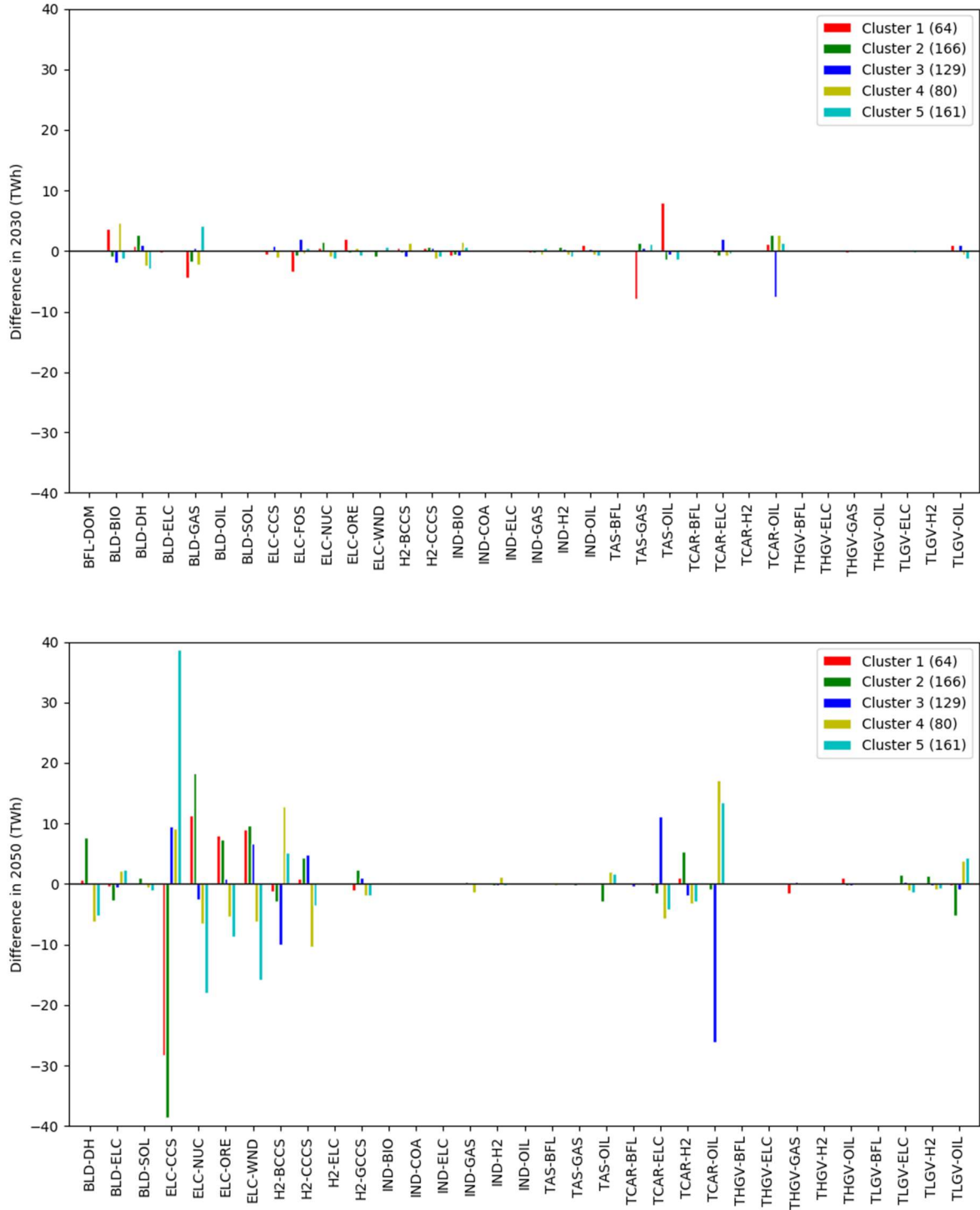


Figure C.4 Differences between the average metrics of all pathways and the mean metrics of clusters in 2030 and 2050 identified with sectoral standardised metrics

Appendix D. Different clustering results of various algorithms

Table D.1 Deviation level of mean metrics of pathways in identified clusters by hierarchical clustering from the corresponding mean metrics of all pathways in 2030 and 2050

Sector		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
		2030	2050	2030	2050	2030	2050	2030	2050	2030	2050
Building	BIO	-		-		+				+	
	DH	+	++		+		+				-
	ELC		-								
	GAS					-					
	SOL										
Electricity	CCS		---		--	-	+++		-		++
	FOS					+					
	NUC	+	+++		++		---		+		--
	ORE		++				--				-
	WND		+++		++		--				-
Hydrogen	BCCS		--		--						+
	CCCS		--		+						-
	GCCS		+++								
Transport	AV	GAS	-								
		OIL	+	-			-				
	CAR	ELC		+		+		+			-
		H2		+++				-			
		OIL		---	-	---	-	--			++
	HGV	GAS						+			
		OIL						-			
	LGV	ELC		+							
		H2									
OIL			--				+			+	

Note: BIO: bioenergy; DH: district heating; ELC: electricity; GAS: gas; SOL: solar thermal; FOS: fossil fuel; NUC: nuclear; ORE: other renewable energy; WND: wind; H2: hydrogen; BCCS: bioenergy+CCS; CCCS: coal+CCS; GCCS: gas+CCS; AV: aviation; CAR: passenger car; HGV: heavy-good vehicle; LGV: light-good vehicle; +/-: positive/negative variation less than 5 TWh; ++/--: positive/negative variation about 10 TWh; +++/---: positive/negative variation about or more than 15 TWh.

Table D.2 Deviation level of mean metrics of pathways in identified clusters by Gaussian mixture model from the corresponding mean metrics of all pathways in 2030 and 2050

Sector		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
		2030	2050	2030	2050	2030	2050	2030	2050	2030	2050
Building	BIO			-		+					
	DH								+		-
	ELC								-		
	GAS					-					
	SOL				+						
Electricity	CCS		+++		-		--		---		++
	FOS										
	NUC		--		+		+		++		--
	ORE		-		+		+		+		-
	WND		-		+		+		++		--
Hydrogen	BCCS				--						++
	CCCS				+				+		-
	GCCS				+				+		
Transport	AV	GAS				-					
		OIL					+		-		
	CAR	ELC				++					-
		H2							+		-
		OIL			-	---		+			++
	HGV	GAS									
		OIL									
	LGV	ELC									
		H2									
OIL					-				-		+

Note: BIO: bioenergy; DH: district heating; ELC: electricity; GAS: gas; SOL: solar thermal; FOS: fossil fuel; NUC: nuclear; ORE: other renewable energy; WND: wind; H2: hydrogen; BCCS: bioenergy+CCS; CCCS: coal+CCS; GCCS: gas+CCS; AV: aviation; CAR: passenger car; HGV: heavy-good vehicle; LGV: light-good vehicle; +/-: positive/negative variation less than 5 TWh; ++/--: positive/negative variation about 10 TWh; +++/---: positive/negative variation about or more than 15 TWh.

Table D.3 Deviation level of mean metrics of pathways in identified clusters by spectral clustering from the corresponding mean metrics of all pathways in 2030 and 2050

Sector		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
		2030	2050	2030	2050	2030	2050	2030	2050	2030	2050
Building	BIO			+		--		--		+	
	DH			-	--		-		-		++
	ELC								+		-
	GAS			-		++		++		-	
	SOL										
Electricity	CCS				+++		++		+++		-
	FOS									+	
	NUC			-	---				---	-	---
	ORE				++		-	-			++
	WND				--	+	-		++		++
Hydrogen	BCCS								+		-
	CCCS				-						
	GCCS										
Transport	AV	GAS									
		OIL									
	CAR	ELC				+				-	+
		H2				-			++		-
		OIL			-	-		+	+		-
	HGV	GAS									
		OIL									
	LGV	ELC									
		H2						+		+	
OIL											

Note: BIO: bioenergy; DH: district heating; ELC: electricity; GAS: gas; SOL: solar thermal; FOS: fossil fuel; NUC: nuclear; ORE: other renewable energy; WND: wind; H2: hydrogen; BCCS: bioenergy+CCS; CCCS: coal+CCS; GCCS: gas+CCS; AV: aviation; CAR: passenger car; HGV: heavy-good vehicle; LGV: light-good vehicle; +/-: positive/negative variation less than 5 TWh; ++/--: positive/negative variation about 10 TWh; +++/---: positive/negative variation about or more than 15 TWh.