527

Using space syntax method to train a model for unsupervised detection of socio-economic conditions - the case of metropolitan area of Tehran

SEPEHR ZHAND & KAYVAN KARIMI,

UCL, LONDON, UK

ABSTRACT

The availability of open-source data, coupled with recent advances in technology has made it easier to create large scale urban and regional models used in the field of environmental studies and specifically space syntax. With the use of open data, large scale regional road-centre line models (Turner, 2005) can now be created and processed to explain the spatial configuration as well as the structure of the built environment. While the study of the socio-economic condition of the built environment correlated with the configuration of the space has been the general use of these models, there has been less focus on multi-layered analysis and metrics across a large model.

On another hand, with restrictions on datasets from formal resources, the conventional use of space syntax theories and methods are limited. However, incorporating more advanced methods of quantitative analysis, space syntax can compensate for the lack of available formal dataset in reading and/or predicting environmental phenomena. Given that with the available data sources such as Open Street Map, consistent spatial network models are available, RCL segment models can be trained to predict the socio-economic condition of areas where the formal data is not obtainable. This research puts forward a workflow through which, the spatial network model can be used to train a model that predicts mentioned phenomena. This workflow uses a large segment model of the metropolitan area of Tehran and uses the centrality measures from space syntax analysis to train an unsupervised model which can predict possible missing information. It also assesses the efficacy of the model and shows to what extent the model is to be trusted and what the shortcomings of the model are.

It is shown that although the models are very efficient in predicting the required conditions there should be a supervised assessment on the parameters of the algorithms to optimize the outcome.

KEYWORDS

Space Syntax, Machine Learning, Unsupervised algorithm, Tehran

1 INTRODUCTION

This paper draws from wider research on the use of open data on explaining and predicting the conditions of urban growth, especially in the context of the developing world. With the fast pace of simultaneous changes in almost every aspect of urban development, it is not easy to understand and explain different aspects of growth and yet it seems more necessary to have a way to do so.

In this respect space syntax provides a method through which some of intrinsic characteristics of the built environment and its changes can be understood given the limited amount of available data. However, the methods developed by the space syntax community has so far been concerned with immediate relations in spatial settings which is used to investigate characteristics of the space. The measures of centrality have been tested against and proven successful in explaining various social, economic, or political properties of the built environment.

Conventionally space syntax models would show correlation between the spatial configuration and other properties of the built environment. This dependency may limit the applicability of the space syntax analysis in certain conditions where there is no formal account of these properties. For example, when researching an evolving urban setting where a considerable part of the change happens through informal dynamics, the datasets for these properties of the study area are either not available or do not exist.

However, approaching the results from the space syntax model can be used to predict certain properties of space even though this data is not directly available. It can be argued that the spatial properties derived from space syntax analysis follow the condition that have evolved before or simultaneous with the changes in the configuration of space. Meaning that the properties of urban configuration can be translated into other properties such as land-use, density etc.

Nevertheless, the use of models and measures suggested by space syntax, has been less the subject of prediction and specifically unsupervised machine learning algorithms which use the results from the space syntax analysis objectively and explain its intrinsic dimensions. While there is a potential for space syntax analysis to be incorporated into the wider platform of data analysis which relies on scientifically endorsed method, and then help the general practitioners of space syntax methods and theories to further examine the relationships that are logically constructed within the built environment.

This would include finding reasonable and reproducible pipelines that could be facilitated by the data science methods and further show the complex relationships that can be explained by the space syntax

2 THEORY

While the study of the built environment through spatial data implies a great extent of possibilities into the ways in which new insight and explanations are at hand, there emerges situations that the in-spite of vast available open datasets, the straightforward methods would not suffice in completing the task. Significantly in the context of developing countries it is the inconsistency in the structure of the datasets, or their coherent availability due to political limitations, which hinder the process.

However, the development in the theories and research that extracts information through simple materiality of the built environment on one hand, and the advancements in aggregating the available data and proven scientific methods to generate models that would predict where the data is not available. This chapter explains the process through which openly available crowd-sourced data can create models that explain a great deal of social, economic, or even political phenomena. Furthermore, it explains in a certain context how and to what extent this model can predict urban conditions through the medium of a general spatial datasets. Accordingly, this chapter explains what the basis of the spatial models are, and to what extent the model can be used to further explore complex urban conditions.

2.1 Open datasets and large spatial models

Recent availability of individual movement data that is captured and stored through hand-held mobile devices as well as the open platforms to add and edit data, has produced mapping possibilities that correspond to the reality of individual and aggregate movement patterns. While these patterns were generally assumed or created, these available datasets not only show patterns but also the way individuals and communities confront and interact. This patterns of movement and behaviour provide a unique chance to examine how they correlate to the process and product of physical growth and development of the built environment, specifically urban space.

the correlation between the built environment and the movement and interaction patterns have long been the subject to research, (Hillier, 2000; Hillier, 1996) and although some might suggest that one happen in the context of the other, there is a broad area of spatial research that suggest the two happen simultaneously, meaning that patterns of movement and collective interaction, and the creation of built space happen incrementally and are impacted by each other (Karimi, 2012). Be that as it may, the availability of these datasets provides an opportunity to examine the characteristics of the built environment as well. Meaning that the movement pattern captured in these datasets would also represent the space available or generated through the movement. In other words, the dataset also represents the spatial arrangement within which everyday life occurs.

This interpretation of this representation of space relies on the way both the movement pattern and the spatial arrangement are correlated. Given that this representation breaks down a large environment into segments that represent available pockets of activity, (Kolovou, et al., 2017)the way these pockets of activity – or convex spaces (Hillier & Hanson, 1984, p. 92) – are connected to each other would facilitate or hinder certain activities and in certain radius of possible movement.

This general representation of the spatial arrangement represented through broken-down pocket convex spaces creates a configuration that puts each of these spaces in relation to others that forms the general urban spatial network. This can be used to assess possibility of a certain activity or behaviour in comparison to the rest of the model. This spatial configuration then not only is shaped by the aggregate of the activities in space, but also becomes the reason for the activity itself. it can be argued that analysing the urban configuration, one can describe the extent to which a certain segment is central to the rest in terms of local or global process. This is important in the sense that understating the interrelations in a settlement follows a dynamical system that is present in larger system that the settlement is emerging in. Thus is a settlement there are patterns that can be traced in the larger system.

3 DATASETS AND METHODS

As stated through, the essential argument in this paper evolves around the ways in which space syntax analysis, and the centrality measures at the heart of it can be relied upon in predicting certain spatial and/or socio-economic conditions of the city. Given that this topic can be discussed through many disciplines, the methodology presented here tries to suggest a workflow, that can be standardized and tested in different conditions.

However, due to the limitations around test data, the applicability of the method is only tested against one large spatial region, while it has been tried to have the standard methods applied as it is validated through their disciplines, specifically the data science methods. These would include generating, cleaning, optimizing, and producing standardized datasets and outputs. Here these methods are explained in order:

3.1 Segment road-centre line model

The analysis in this research uses a combination of models that would capture the spatial characteristics of the space, which is then used to make the predictions. Initially a road-centre line model is created through using OpenStreetMap data, which shows the lines of movement in a spatial setting and represents these movement paths as lines. This dataset contains the information from the crowd-sourced dataset that are constantly updated and revised and corresponds to the paths that have been solidified by frequent vehicular movement.

The specific dataset used for the model is downloaded from classified Geofabrik download servers (Anon., 2019) which is then simplified and segmented. This process of simplification and segmentation would create a model that is drawn from the idea of axial lines map in representing the space (Hillier & Hanson, 1984, p. 92) as it maximizes the axial movement within one continuous convex space. As research shows there is encouraging results comparing Axial and segment maps (Kolovou, et al., 2017), the practicality of RCL lines in using wed mapping services and voluntary datasets would provide the chance to create larger models.

Given the methodological approach in this research, an RCL map of the metropolitan area of Tehran was created which covers not only the legal boundary of the province, but also all the satellite and marginal settlements, that in a wider context are dependent to the urban area of Tehran, and their dwellers commute to the city on a daily basis (Ghomami, et al., 2001).
This model was then cleaned and simplified, to an extent where between every junction there is only one segment representing the space, and certain topological errors such as parallel lines, angular changes, and unnecessary details were removed (Krenz, 2017).

3.2 Space syntax analysis and centrality measures

Understanding aggregate behaviour and accordingly the societal relations in a large setting such as a metropolitan area can become dramatically hard. While there has been a debate between configuration and attraction as rival concepts in accounting for urban movement, Hillier suggests that the configuration of the urban grid would have an impact on generation of centres that are economically engaged in the larger economic processes of a city and a region. (Hillier, 2000). Through this understating the local and global rules (Hillier & Hanson, 1984, p. 201) that govern the configuration and evolution of space can explain the emergence and change in the surface fabric of the city.

Therefore for understating the functionality of the configuration in this large system a set of centrality measures such as closeness centrality (Sabidussi, 1966) and betweenness centrality (Freeman, 1977) were employed to assess the local and global depth and integration of the segments in relation to the overall configuration. This model consisting of almost 700k segments was analysed using DepthmapX 0.5 (Turner, et al., n.d.) and the centrality measures were derived from local radius r400m to global radius.

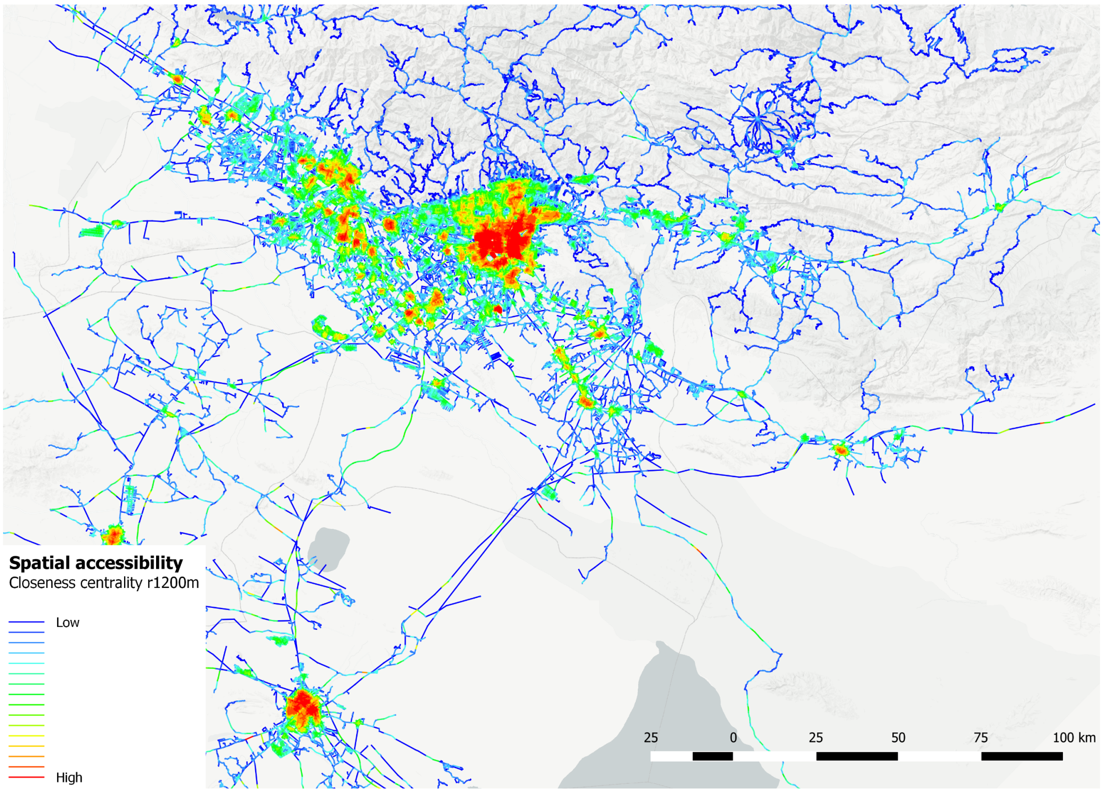


Figure 1 RCL model of the metropolitan area of Tehran analysed for centrality measures. This map shows the closeness centrality (Integration) at 1200m which shows the dense centres of all the satellite cities with warmer colours

3.3 Plot shapefiles for Tehran and satellite cities

Furthermore, a set of GIS data files consisting of some spatial characteristics of space, from official sources were incorporated into the model to help assess the validity of the analysis. This includes a shapefile from the Municipality of Tehran consisting of plot-based data with building land use data as well as plot area and perimeter. (Municipality of Tehran, 2018).

The same datasets from the New Towns Development Company (NTDC) that is responsible for planning and upgrading strategies in new towns and informal settlements. This includes a range of cities varying in type:

- Salehieh: an old village that has been developed into a new planned city
- Ferdowsieh: An organically grown village
- Malard: a small ancient city with an expansion of informal subdivided land
- Arjomand: an organic city with recent structural readjustment

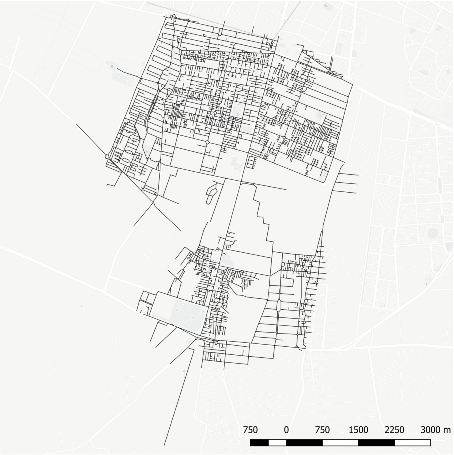


Figure 2 Malard, organic city with informal land subdivision

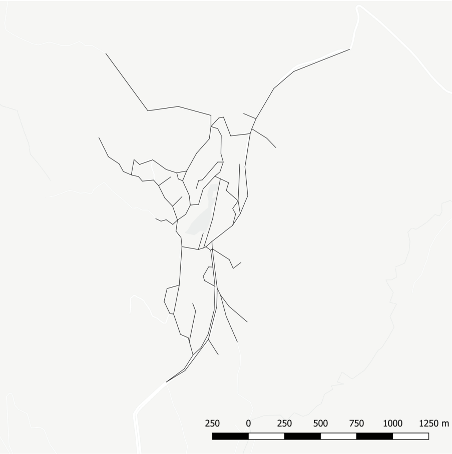


Figure 3 Arjomand, an organic settlement

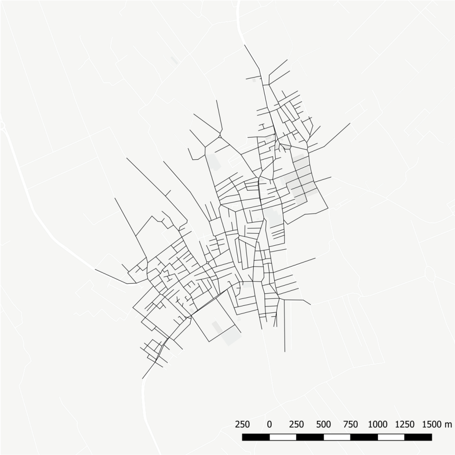


Figure 4 Ferdowsieh, a medium sized city of organic structure

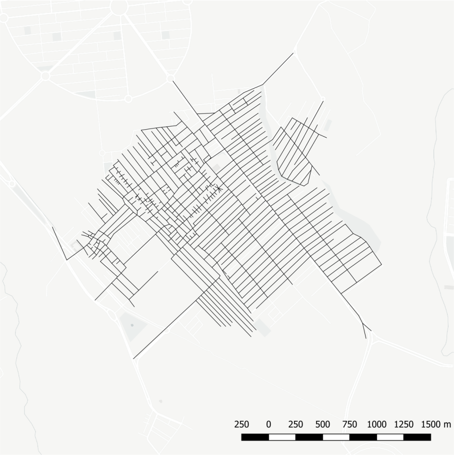


Figure 5 Salehieh' a new planned settlement

These cities that vary in size and typology are all part of the larger spatial network model and are selected to be tested against the more comprehensive model of Tehran.

3.4 Spatial data set and dimensionality reduction

Aiming to capture all the possible spatial characteristics of the region, the syntax analysis in this research ranges from local radius of 400m to the global radii of 40km and rN. The result of this analysis would create a dataset of more than 40 labels with over 700k features. And with the limited computational power, complex analysis of this dataset would be practically impossible. While having to maintain all the spatial characteristics of the space without removing or reducing any of the data labels, a method of dimensionality reduction was employed to reduce the number of dataset labels while accounting for all the measures and radii.

The reduction of these components while keeping the properties of the space was done using the *Principal Component Analysis* which would read through all the labels of the dataset and reproduce a smaller number of components that would partially correlate to the original labels. This process can be described as follows:

Given a dataset of ℓ features each represented by p-dimensional vectors of wights $\mathbf{W}_{(k)} =$ $= \left(w_1, \dots, w_p\right)_{(k)}$, the PCA will transform the dataset in a way that each row $\boldsymbol{x}_i$ of the new dataset $\mathbf{X}$ has new weights $\boldsymbol{t}_{(i)} = (t_1, \dots, t_\ell)_{(i)}$ where:

$$t_{k_{(i)}} = \boldsymbol{x}_{(i)}.\mathbf{W}_{(k)} \; for \; i = 1, \dots, n \; and \; k = 1, \dots \ell$$

This is done in a way that each of the new variables $t_1, \dots, t_\ell$ of $\boldsymbol{t}$ throughout the dataset, successfully inherits the maximum possible variance from the original set. This will reduce the dimension of the dataset while maintaining the characteristics of the dataset and reducing the computational complexity of further analysis.

In this project the principal component analysis (PCA) from the Scikit learn libraries (Pedregosa, et al., 2011) was used to reduce the number of components while merging the data from all the labels. This process was repeated for both the Normalized values and non-normalized values, and it seems that the algorithm better explains the variance ratio in non-normalized values as opposed to the normalized ones.

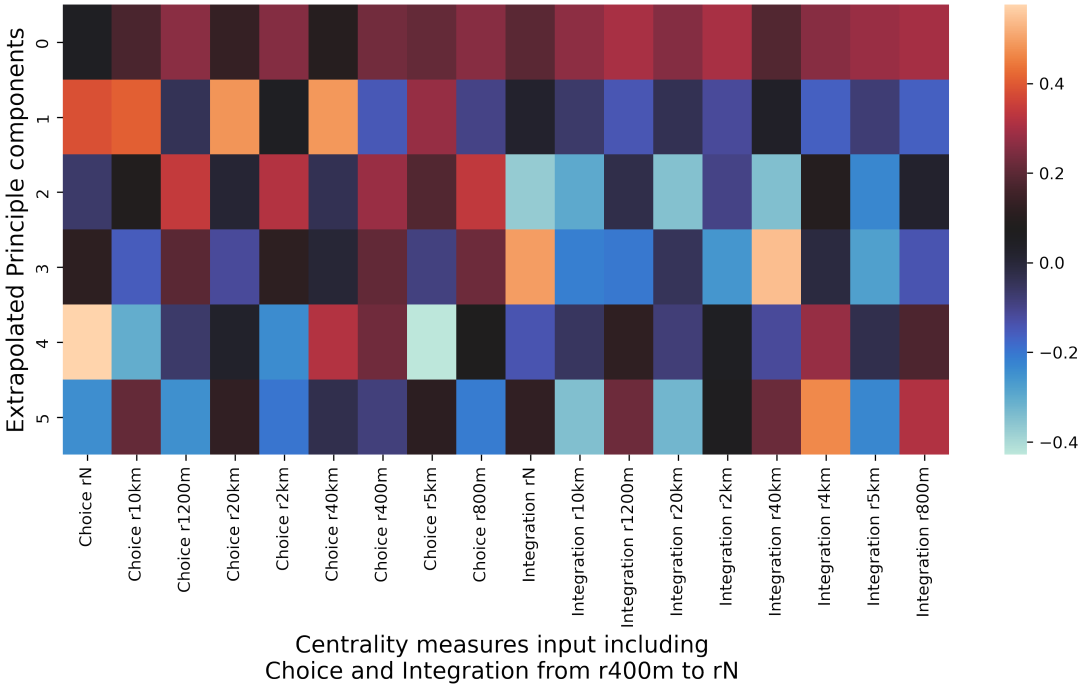


Figure 6 Explained variance ratio by each of the 6 components while reducing the non-normalized values of the dataset from 17 to 6

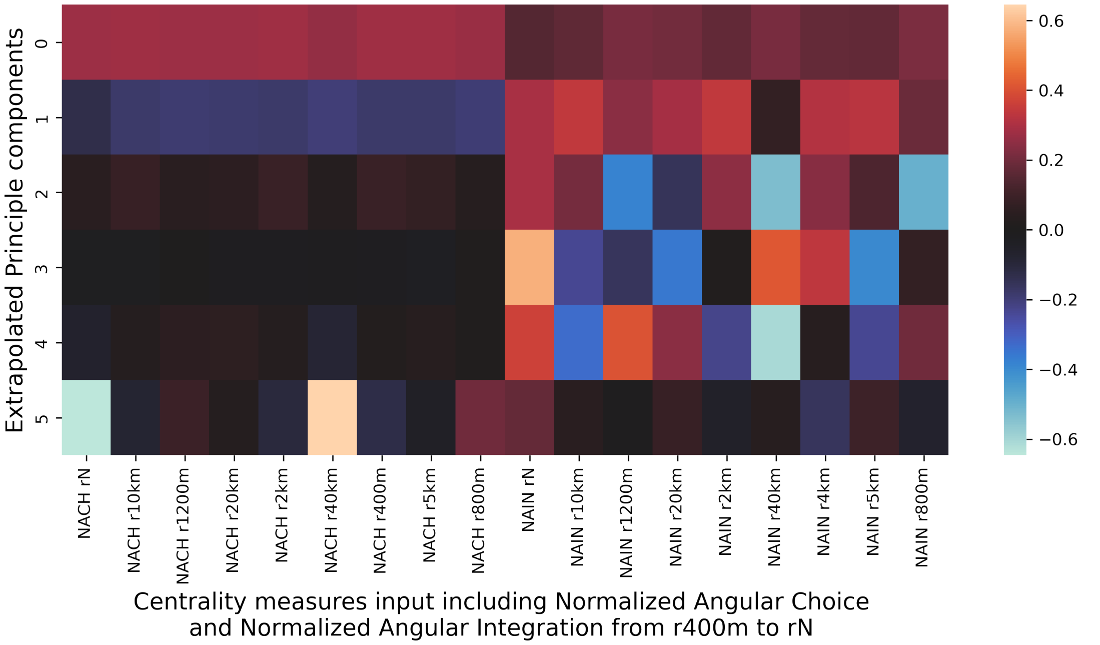


Figure 7 Explained variance ratio by 6 components while reducing the normalized values to 6 components

The PCA analysis applied to the normalized values does not produce results that correlate to the original values, specifically in terms of betweenness centrality. The explained variance ratio in fig 6 shows a higher correlation with all the input values across all components, while in fig 7 half the input values – corresponding to the choice measure in different radii – does not have strong correlation with the components. This resulted in preparing the training, and test sets using only the non-normalized values from the syntax analysis.

3.5 Clustering, classification, and model training

The theoretical approach in the research evolving around the correlation between generation and evolution of the built environment and the element of movement (Hillier, 2000) as well as the modelling and analysis approach are all set to identify the intrinsic characteristics of the urban configuration that are generated through time.

This suggests that within the continuous spatial network that functions with the same economic driver, the same spatial characteristics should appear if the same local and global rules are present. In other words, if the analysis seems to simultaneously detect similar values e.g., local integration, global choice, and segment length, then there is high probability that the segments that correspond to these values, should have the same interrelations in the spatial network. Given Hillier's discussion on *centrality as a process,* (Hillier, 2000) segments that have the same relations and correspond to similar local and global rules are more likely to accommodate the same spatial characteristics as well.

In this way capturing the similarities of the segments that are close in the mentioned characteristics would introduce the possibility of type in the built environment that could be detected through its relations in a network. Thus, in this research K-means clustering algorithm was used to detect the segments that are closer in terms of their characteristics. This algorithm can be scripted as:

$$\arg\min \sum_{i=1}^{k} \sum_{x \in S_i} \| x - \mu_i \|^2$$

Given a set of data points (x_1, x_2,…,x_n) where each of these data points have d-dimensions (in this segments are the datapoints and the centrality measures are the dimensions), the algorithm aims to partition the set of data points into k ($\leq n$) clusters (**S** = {S_1, S_2, … S_k}) in a way that within each cluster the sum squares (variance) is minimized. Here μ_i is the average points in the S_i and the formula repeats until the variance is minimized.

Given the specificities of this research, the centrality measures are understood as the intrinsic properties of the space and applying the clustering process would group the segments together that share the most similarities in terms of their relationship to the network.

The number of clusters and the number of components in this process is determined through calculating the silhouette score for different combination of PCA outputs. In this process the silhouette coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. (Pedregosa, et al., 2011). This process can be summarized in the following formula:

$$\boldsymbol{s}_{(i)} = \frac{b_{(i)} - a_{(i)}}{\max \{a_{(i)}, b_{(i)}\}}, if\ |C_I| > 1$$

Where C_I is the number of datapoints assigned to cluster $\boldsymbol{i}$**.** The silhouette score (or the clustering coefficient) generally demonstrates how the dataset can be divided into clusters and shows how distinct these clusters are. Thus, the higher the silhouette score the more distinguishable the clusters. However, in this research and generally in data science this does not ensure better results in the overall research given that the 2 cluster would usually have higher silhouette score than higher numbers.

Having applied the PCA analysis on the non-normalized centrality measure datasets, the original dataset was transformed into new datasets reducing from 18 to a range from 2 to 10 components. A silhouette score analysis was done on each of the new datasets which were each clustered between 2 to 18 different clusters. Plotting these results show that there is a meaningful change in the silhouette score for all the PCA results when the number of clusters they are clustered to is lower than 5.

In data science using this break point to determine the optimum number of clusters is called the elbow method (Thorndike, 1953). Given that each silhouette score shows the explained variance by each of the clusters, the clusters before this point have low number of clusters with high clustering coefficient, and after this break point have relatively similar clustering coefficient. Therefore, this process shows that a 6-cluster configuration would identify the most distinct segments.

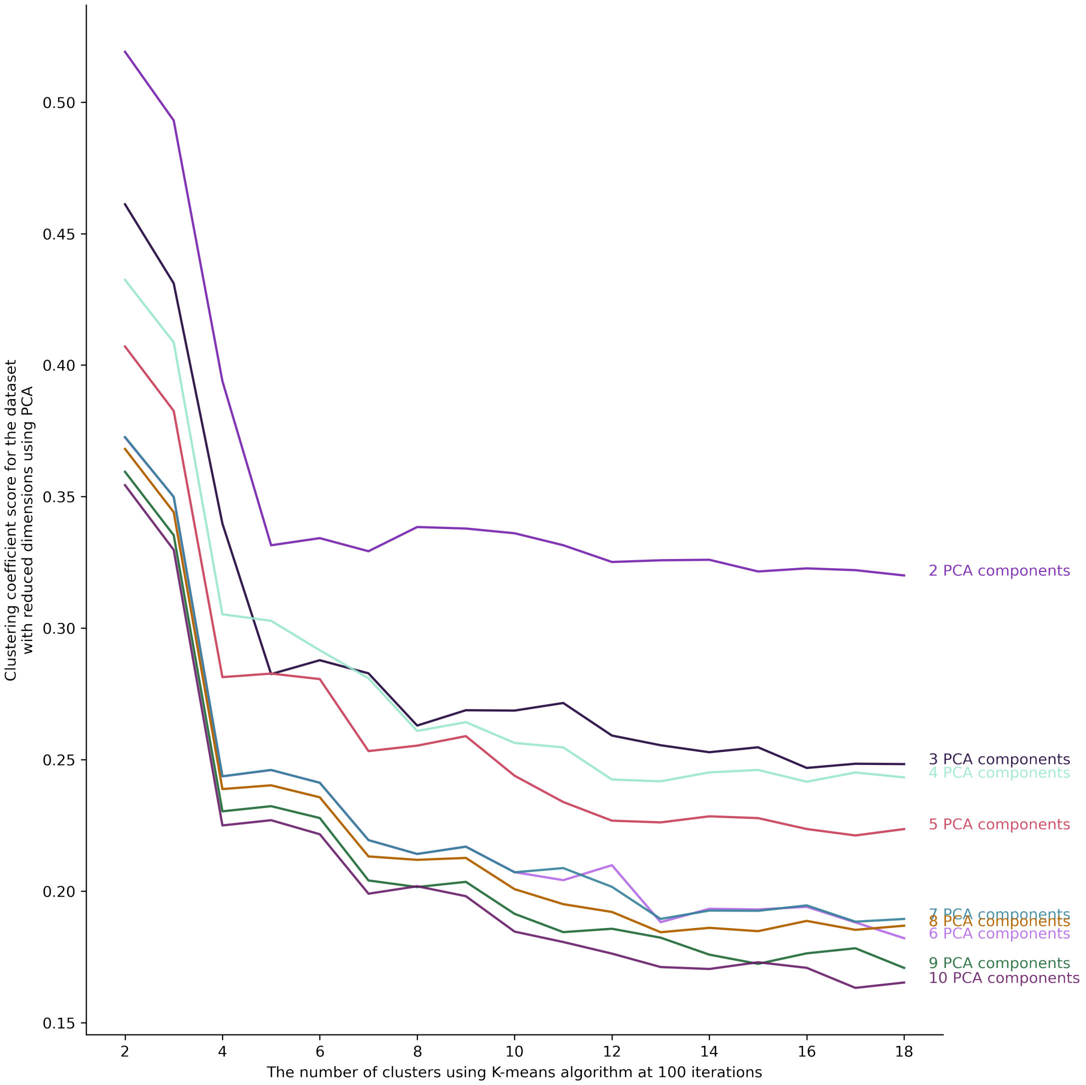


Figure 8 Silhouette score (clustering coefficient) for the rising number of clusters for different number of components from PCA analysis

Upon determining the right number of clusters, the spatial network is joined with the official dataset with land use and area data, and clustered into 6 different clusters using K-means algorithm (MacQueen, 1967) The result of this process (cropped to the extents of the legal boundaries of the city of Tehran as the training dataset) is used to train a KNN classifier (K nearest neighbour).

Although out of the discussion in this paper, the initial result from the clustering analysis that incorporates all the centrality measures and is done on the optimum number of clusters, distinguishes the areas in the current city that have been developed under fundamentally different criteria and have distinctly diverse characters. Highlighted in different colours in figure 9, each cluster belongs to a different era of development where the approach to city planning was different.

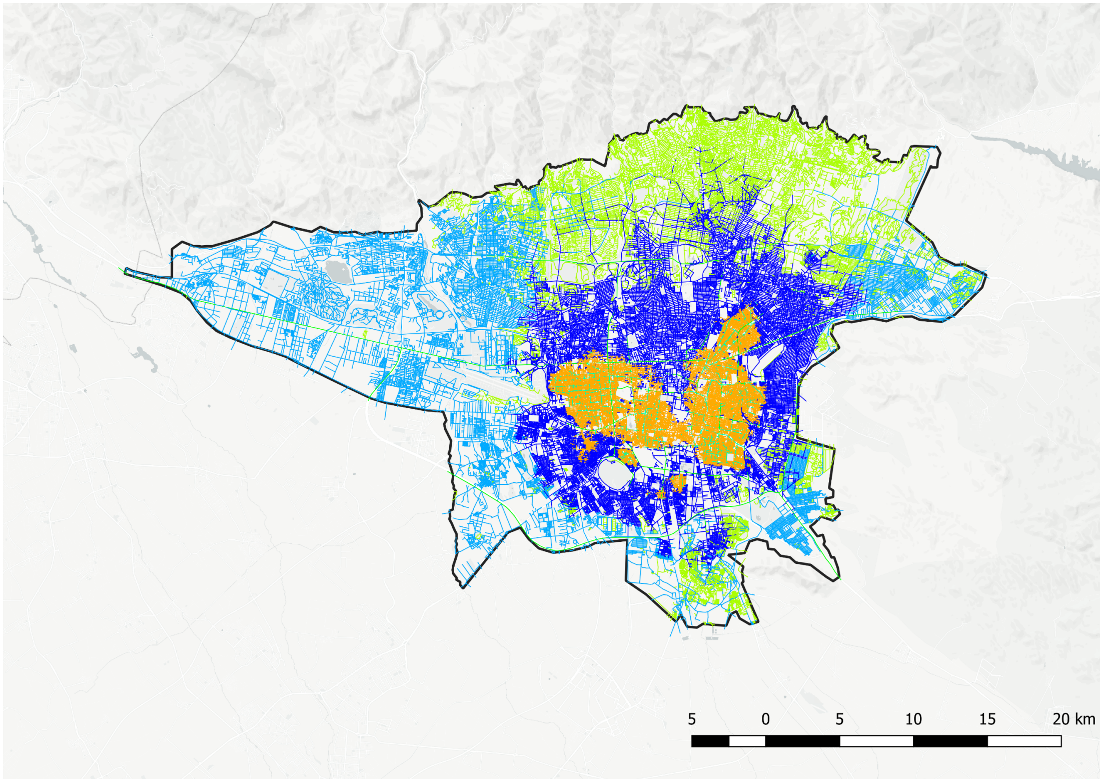


Figure 9 The city of Tehran classified into 6 different patterns that correspond all the centrality measures. each colour corresponds to tone cluster

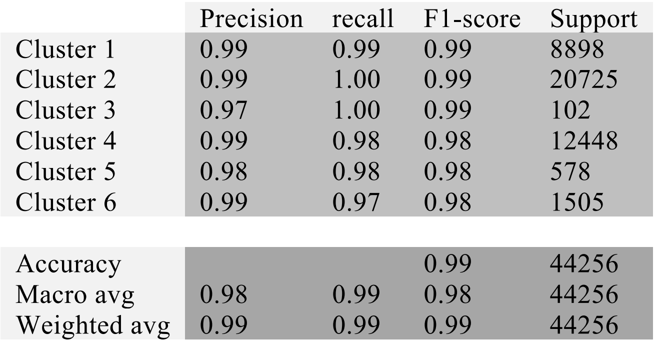
With each segment assigned to a certain cluster, this dataset is set against the new settlements' dataset to determine which segments of the new settlements are closer in terms of their centrality properties. This is done through KNN classifier that implements the *Minkowski distance* function. This algorithm then calculates the metric distance between each point and trained model (in this case the spatial network model of Tehran) and determines which cluster each segment belongs to. This function can be summarized as:

$$D(X,Y) = \frac{\sum_{i=1}^{n} |x_i - y_i|^p}{p}$$

Where $X = (x_1, x_2, \ldots, x_n)$ *and* $Y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$ are the two points from the order of $\mathcal{p}$ that the function calculates the distance between. This trained model is used to identify the cluster that each segment in each of the other samples are similar to, based on merely their spatial values. Based on this the correlation between other morphological characteristics of the training set (the city of Tehran) and the 4 cities can show to what extent the spatial model can be relied upon in accounting for morphological characteristics.

4 RESULTS

The application of data science methods and unsupervised learning algorithm to the dataset created through centrality measure analysis does reveal some interesting results, however the interpretation, as well as comparing the results applied to the unsupervised learning mechanisms would require some further human insight. Nevertheless, the metrics used to assess machine learning models show that the model trained in this research was able to predict the clusters of other segments with high accuracy.



	Precision	recall	F1-score	Support
Cluster 1	0.99	0.99	0.99	8898
Cluster 2	0.99	1.00	0.99	20725
Cluster 3	0.97	1.00	0.99	102
Cluster 4	0.99	0.98	0.98	12448
Cluster 5	0.98	0.98	0.98	578
Cluster 6	0.99	0.97	0.98	1505
Accuracy			0.99	44256
Macro avg	0.98	0.99	0.98	44256
Weighted avg	0.99	0.99	0.99	44256

Figure 10 Accuracy assessment of the model based on a training set split to 0.7 training set and 0.3 test set

While fitting the model and testing it against the 4 other cities returns results with highest possible accuracy the interpretation of the results as well as the extent to which the model can predict other spatial and morphological characteristics requires further investigation.

Validating the results from the prediction, was set against the available data from the training set to see if the unsupervised assignment of the cluster for each segment could have predicted other morphological characteristics with high accuracy. Due to the limitation of this study it only looked into the way in which the model is able to predict the segments in terms the average plot size and dominated land-use.

With different sizes of the datasets, this comparison was done using a standard normalization. In the case of the land-use comparison, land-use of the plots were aggregated onto the closest segment and counted to determine the dominating land-use in terms of abundance. The results from this comparison show a close pattern in terms of the way the predicted segments are characterized by the land-use.

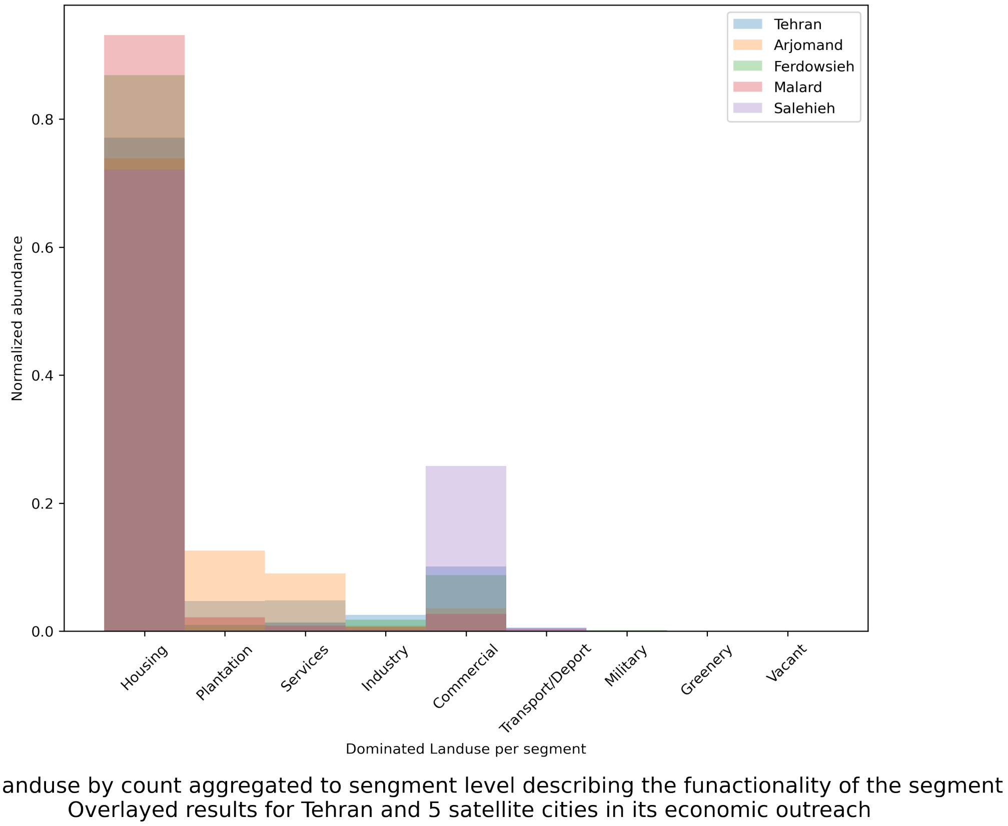


Figure 11 Distribution of the most frequent land-use per segment in Tehran (as the training set) and four satellite cities.

The above histogram shows the abundance of different land-use of 5 cities (Tehran and 4 satellite cities). While overlayed this analysis shows that there is a similar pattern in the way in which different land uses have dominated the segment, which has been repeated in all the settlements. Except for Salehieh which has been predicted to have more commercial land uses and Arjomand to have more plantation and services, the same pattern of land use seems to be repeated in the predicted segments.

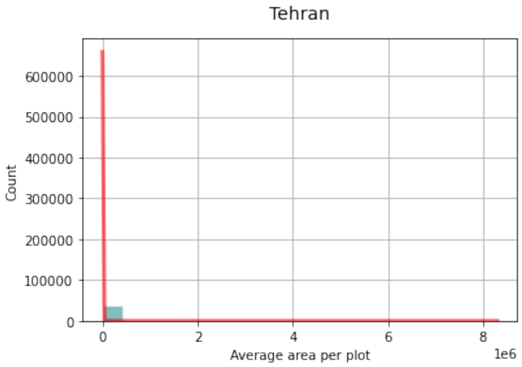


Figure 12 Distribution of the average parcel area per segment in the training set of Tehran with the same clusters as in the other 4 cities

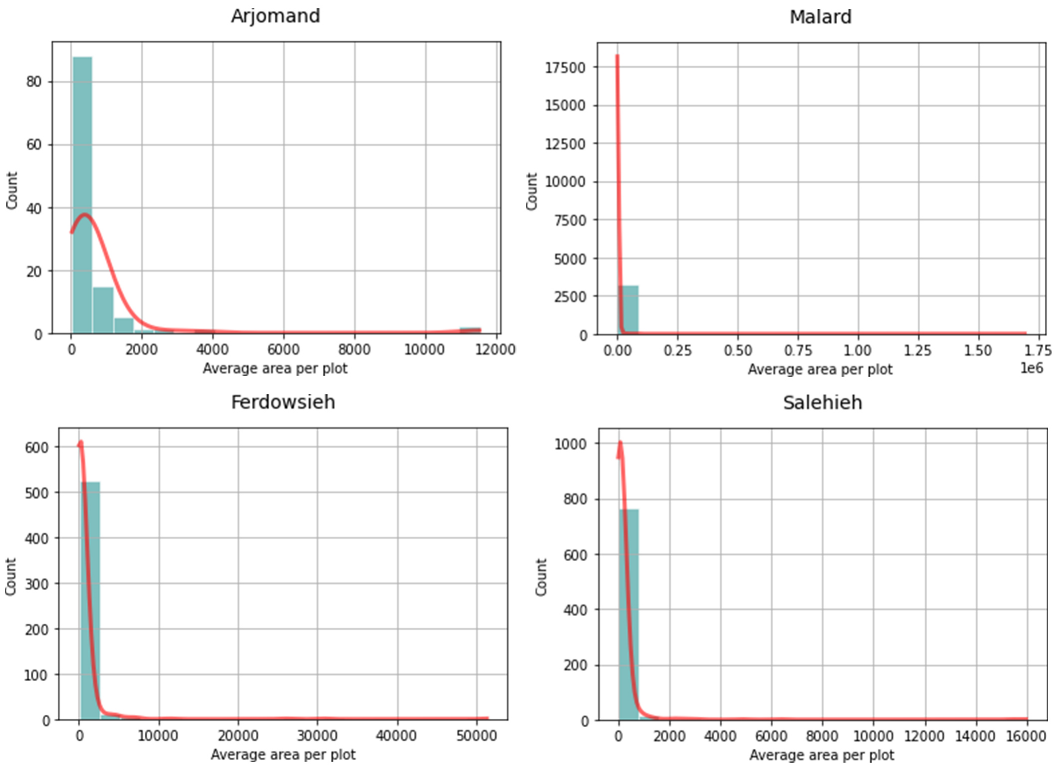


Figure 13 Distribution of average parcel area per segments in the 4 cities

	Tehran	Arjomand	Ferdowsieh	Salehieh	Malard
Mean area per segment	1990.15	693.93	527.78	202.15	1120.99
Median area value	67.99	380.63	127.18	85.43	115.46

These summary of these results shows that while there are similarities in distribution of average plot area per segments in the clusters that were picked up by the model, there is more similirities between Malard and Tehran, rather than the smaller cities of Arjomand, Ferdowsieh and Salehieh.

5 CONCLUSIONS

This paper tries to investigate how and to what extent the space syntax analysis can be relied upon to be used in more advanced data science methods in areas of research and practice that the absence of data hinders the work. Applying an unsupervised learning algorithm to a large model, created and analysed through space syntax methods and theories, this research tries to show that while the morphological characteristics of space are captured in the spatial network model, predicting them solely based on these models may not be as straight forward.

Given the standard process in the data science disciplines that is designed to optimize the output of such models, there seems to be still a lot required from a trained human perspective to put these tools to the best use. Specific to this research the number of clusters that were suggested through silhouette score analysis, provide a very general classification for urban type that although meets the standard processing metrics, it does produce encouragingly accurate results.

The research presented here was able to predict spatial characteristics in areas that were closer in size to the training set and with an optimized number of clusters this could be further improved. Thus, the next step for this research would be to test the data using another number of clusters that entails more detailed characteristics and testing it against more morphological metrics for better understanding of the process.

6 REFERENCES

Anon., 2019. *Geofabrik download server.* [Online]
Available at: https://download.geofabrik.de/asia/iran.html
[Accessed 06 08 2019].

Freeman, L. C., 1977. A Set of Measures of Centrality Based on Betweenness. *Sociometry,* Volume 40, pp. 35-41.

Ghomami, M. et al., 2001. *The Plan for the Metropolitan Area of Tehran and Its Sattelite Cities,* Tehran: Centre for research and studies on architecture and cities.

Hillier, B., 1996. Cities as movement economies. *URBAN DESIGN international,* 1(1), pp. 41-60.

Hillier, B., 2000. Centrality as a process: accounting for attraction inequalities in deformed grids. *Urban Design International,* Volume 3/4, pp. 107-127.

Hillier, B. & Hanson, J., 1984. *The Social Logic of Space.* Cambridge: Cambridge University Press.

Karimi, K., 2012. A Configurational Approach to Analytical Urban Design: ‘Space Syntax’ Methodology.. *URBAN DESIGN International,* 17(4), pp. 297-318.

Kolovou, I. et al., 2017. *Road centre line simplification principles for angular segment analysis.* Lisbon, Instituto Superior Técnico.

Krenz, K., 2017. *EMPLOYING VOLUNTEERED GEOGRAPHIC INFORMATION IN SPACE SYNTAX ANALYSIS.* Lisbon, Lisbon Instituto Superior Técnico, Departamento de Engenharia Civil, Arquitetura e Georrecursos, pp. 150.1-150.26.

MacQueen, J., 1967. *Some methods for classification and analysis of multivariate observations.* Berkeley, University of California Press, pp. 281-297.

Municipality of Tehran, 2018. *Greater Tehran Land use data,* Tehran : Municipality of Tehran.

Pedregosa, F. et al., 2011. Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research,* Volume 12, pp. 2825-2830.

Sabidussi, G., 1966. The centrality index of a graph. *Psychometrika,* Volume 31, pp. 581-603.

Thorndike, R. L., 1953. Who belongs in the family. *Psychometrika,* Volume 18, pp. 267-276.

Turner, A., 2005. *Could A Road-centre Line Be An Axial Line In Disguise?.* Delft, TU Delft.

Turner, A. et al., n.d. *depthmapX - multi-platform spatial network analyses software.* [Online]
Available at: https://github.com/SpaceGroupUCL/depthmapX
[Accessed 20 07 2020].