Unravelling the Dynamics of Urban Demographics

Darsh Modi, *Student, TU Delft,*

***Abstract*—In the Netherlands, and in Western Europe overall, residential segregation of ethnic groups is undesirable, and is a growing concern for local authorities and policy makers [1]. Capturing trends in demographic change leading to residential segregation and identifying their causes can help understand the issue. In this project, areas undergoing similar demographic evolution in the city of Delft between 2015 and 2020 will be identified and their causes will be researched.**

***Index Terms*—Residential Segregation, Demographic Dynamics, Hierarchical Clustering, Unsupervised Learning**

I. Introduction

CITIES are mixing grounds that bring together diverse social factions based on age, education, and ethnicity, thus holding the promise of social integration of people from all walks of life. However, the spatial segregation of these ethnic and economic groups can impede interaction and be a hurdle to the integration process [2]. Investigating the influence of urban planning on the demographic dynamics within neighborhoods can empower urban planners to design environments conducive to social integration. This research aims to use unsupervised machine learning methods to visualize demographic data from various neighborhoods in the Netherlands, namely ones in the city of Delft, from 2015 to 2020, to see how they change. The objective is to categorize neighborhoods into clusters with similar demographic shifts and examine how these shifts correspond to urban planning and development.

II. Methods

Data obtained from the Dutch office of statistics (Centraal Bureau voor de Statistiek, Open Data) has been analyzed using Python (Pandas [3], GeoPandas [4], MatplotLib [5]), QGIS [6], and Orange Data Mining [7]. Postcode level information about various attributes such as gender mix, household composition, nationality distribution, percentage of people from Western and non-Western backgrounds, etc. across time (2015 - 2020) was available in this data. Their moving averages were calculated and unsupervised learning, specifically hierarchical clustering, was employed on the difference between 2015 and 2020 numbers to cluster groups of similar areas together to identify trends in the data. Google Colab was used to write and run the code.

A. Hierarchical Clustering

The project started on an exploratory foot, with clustering being the initial method of choice, with a possibility of using other methods as progress was made. Hierarchical clustering was selected amongst other methods.

D. Modi is a student at the department of Technology, Policy, and Management, Delft Institute of Technology, Delft, 2628BX, Netherlands

1) Hyperparameters: The methodology employed involved careful consideration of hyperparameters, particularly the number of clusters, which were not constrained to be continuous in space. This implies that while clusters may share common characteristics, their spatial distribution was not a factor in their formation. The goal was to determine the smallest number of clusters that could still reflect significant differences across clusters, while ensuring homogeneity within each cluster. The process began with two clusters and expanded until no substantial differences were observed at five clusters.

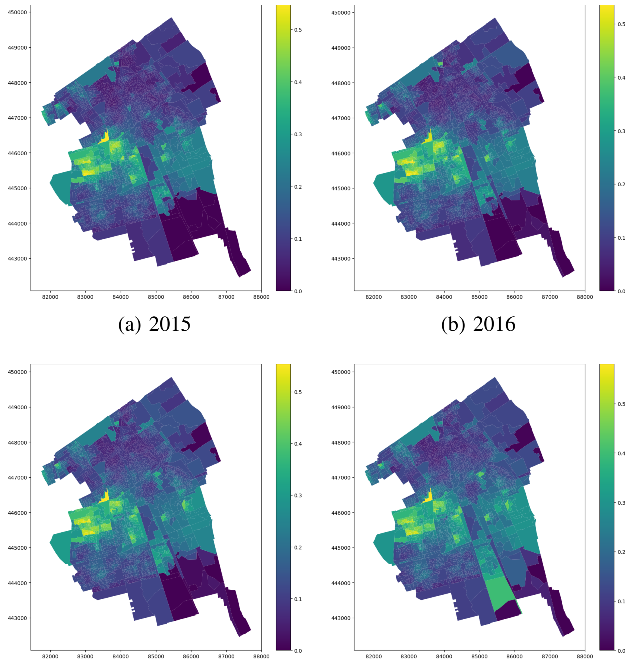
Ward's method was the linkage method of choice. This method was used as it aims to minimize variance within clusters, leading to a high degree of homogeneity — a key objective of the study. Ward's method is particularly relevant when the focus is on the average of the clusters, with all observations gravitating around that mean, providing a clear and interpretable grouping.

2) Other Clustering Algorithms: The study also considered other clustering methods. For instance, K-means clustering shares similarities with hierarchical clustering when Ward's distance is used, as both aim for homogeneity within clusters. However, hierarchical clustering was preferred for its deterministic nature, as opposed to the potential randomness in the initial seeding of K-means 9. DBSCAN, another clustering technique, is generally regarded as effective but was found to be less suitable for this specific dataset, failing to distinctly identify clusters 8. This selection process highlighted the importance of matching the clustering technique to the characteristics of the data to ensure the most coherent and meaningful categorization. The library used to implement clustering and for data preprocessing was scikit-learn [8], [9].

III. Results

Firstly, the evolution of demographics was observed through the change in share of people from non-Western migration backgrounds in a given area (Fig. 1).

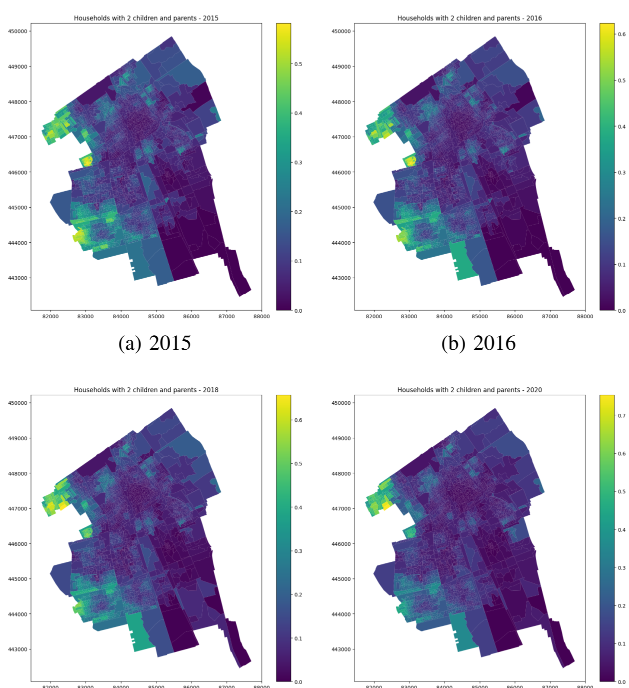
Household composition was also an interesting variable, as it showed the change in types of people moving into and out of Delft (Fig. 2).



(a) 2015 (b) 2016

(c) 2018 (d) 2020

Fig. 1. Share of immigrants from a non-Western background in a given postcode over the years. The scale on the right denotes the proportion of immigrants from non-Western backgrounds.



(a) 2015 (b) 2016

(c) 2018 (d) 2020

Fig. 2. Households with 2 children and parents in a given postcode over the years. The scale on the right denotes the proportion of this demographic.

The change in these variables is clearly visible. Next, a hierarchical clustering was performed using variables that denote the share of -

1) Immigrants from non-Western backgrounds (NWB)
2) Immigrants from Western backgrounds (WB)
3) Dutch nationals (NL)
4) Households with 1 person (HH1)
5) Households with 2 or more people, no children (HH2+NC)
6) Households with 2 or more people, with children, with 1 parent (HH2+CSP)
7) Households with 2 or more people, with children, with 2 parents (HH2+C2P)

This is what the algorithm yielded -

TABLE I
CLUSTER COMPOSITIONS - % CHANGE IN VALUE

	NWB	NL	WB	HH1	HH2+CSP	HH2+C2P
0	0.8	-1.7	-0.2	-3.1	2.5	-0.6
1	2.9	-5.8	3.4	-0.5	-0.2	0.0
2	4.9	-7.4	3.1	5.2	-0.1	-1.3
3	-2.2	3.4	3.6	-13.8	-5.7	16.8
4	-22.5	12.3	-14.8	2.4	0.4	0.4
5	1.9	-3.1	0.8	1.6	0.1	-3.4

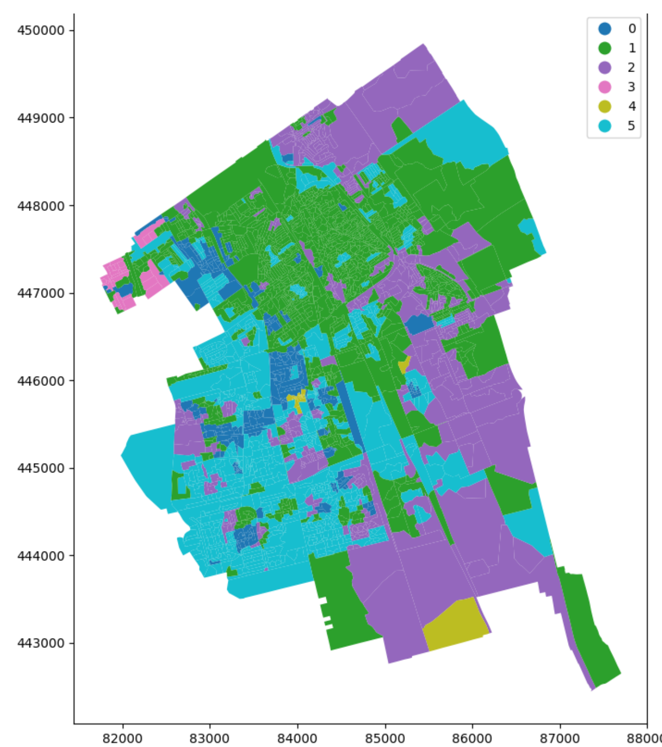


Fig. 3. Result from Hierarchical Clustering - The City of Delft

TABLE II
MEANS OF VALUES OF VARIABLES USED FOR CLUSTERING

NWB	NL	WB	HH1	HH2+CSP	HH2+C2P
2.38	-4.59	2.04	0.62	0.16	-1.02

2 clusters are attention grabbing (by comparing cluster values to mean values for those variables), namely clusters 3 and 4.

A. Cluster 3

Using street view on Google Maps, the following images were obtained -

Fig. 4. Location 1 in 2010

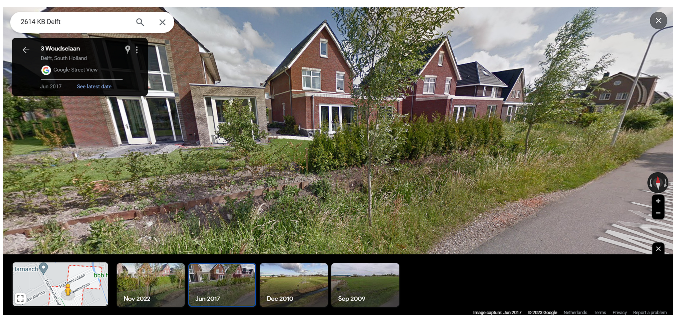


Fig. 5. Location 1 in 2017

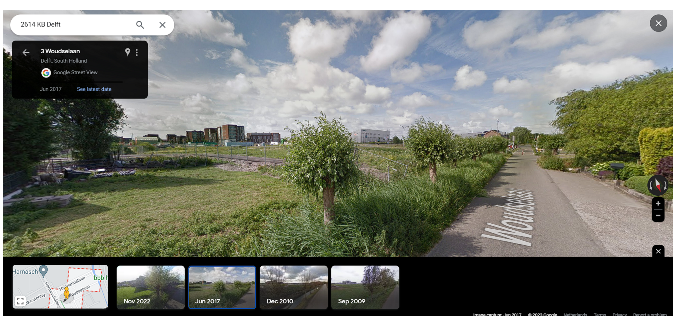


Fig. 6. Location 2 in 2017

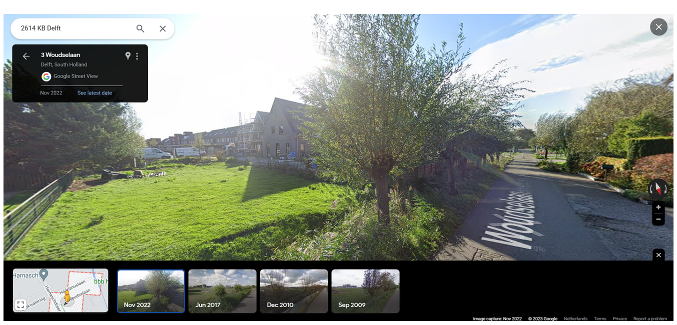


Fig. 7. Location 2 in 2022

IV. Discussion

A. Cluster 3 - Homes

Located on the North-West of the city, it can be seen that there is an exodus of one person households and an influx of families (2 children, 2 parents). From Google Maps, the corresponding list of postcodes belonging to that cluster revealed that there have been lots of new homes built there, as seen in figures 4 through 7 ([10], [11]). This is an example of how infrastructure development has lead to a change in the demographics of an area.

B. Cluster 4 - Student Housing

Cluster 4 consists of student housing (Roland Holstlaan 1-749 and Professor Schermerhornstraat 9-127) managed by the Dutch company DUWO. They offer student housing in the Netherlands. TU Delft has a partnership with DUWO for incoming students every year, to whom they provide limited duration stays. From the data, there has been an increase in the number of Dutch citizens in these student accommodations, and a decline in the number of international students in the buildings. DUWO only reserves rooms in specific buildings (Van Hasseltlaan, Korvezeestraat, and Röntgenweg) for international students [12], and to keep it as fair as possible for everyone, the rooms that become available in these buildings are alternately advertised to Dutch students and to international students through ROOM.NL, which gives you preference based on your date of registration (older is better) [13]. Policies such as this influence the mix of residents that live in an area, and they can be used to influence the demographics of a part of the city.

C. Other Methods

Another possibility would be to use the principal component analysis technique to discover trends in the data, as a part of further exploration to see what results come out of it. Due to the limited time available for the project, and the success with hierarchical clustering, this technique was not explored.

D. The City of Delft

The city of Delft is a city the author is most familiar with in the Netherlands, and hence it was chosen for this project. It is also smaller in size compared to Rotterdam/Amsterdam, allowing for ease of computing, given the limited amount of resources.

V. Conclusion

Infrastructure development in the city of Delft in the form of houses, and policies of student housing are seen to have had an effect on the demographics of the people who live there. This project demonstrates a proof of this very concept. This study can be extended to other cities in the Netherlands to identify if similar infrastructure development and policies have the same effect across the country, and possibly across comparable countries as well. Using this as a tool, governments and policy makers can plan projects using data to achieve a

desired demographic mix to develop themselves as desired, economically and socially.
Although this tool could be useful for urban planning, it has the potential to cause the gentrification of neighbourhoods by displacing lower income groups and reducing the availability of affordable housing, which is already a big problem in the Netherlands. Relevant stakeholders must be involved in this process to ensure equitable and sustainable development going forward, if this kind of data driven approach is used.

Acknowledgment

The author would like to thank Lucas Spierenburg, a PhD student at TU Delft. He came up with the idea of the project and was a guiding force throughout the process, without whom this would not have been possible. The author would also like to thank the team behind the Fundamentals of AI Program offered by the EEMCS faculty at TU Delft, through which this project was undertaken, and by whom the knowledge required to carry it through was disseminated.

Appendix A
Code for the Project

The code for this project was written in Python, in a Google Colab notebook, which is publicly accessible here - Notebook. Data can be found here - Data.

Appendix B
Output from Other Algorithms

A. DBSCAN

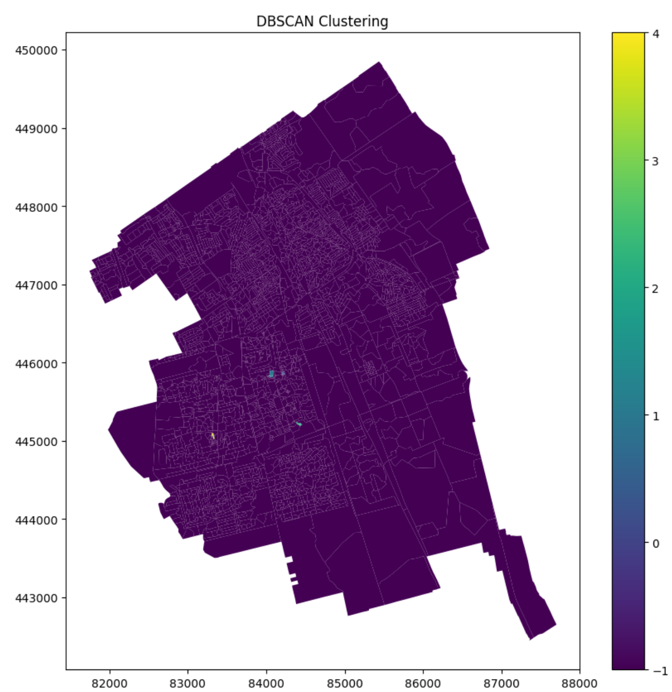


Fig. 8. Result from DBSCAN Clustering - The City of Delft

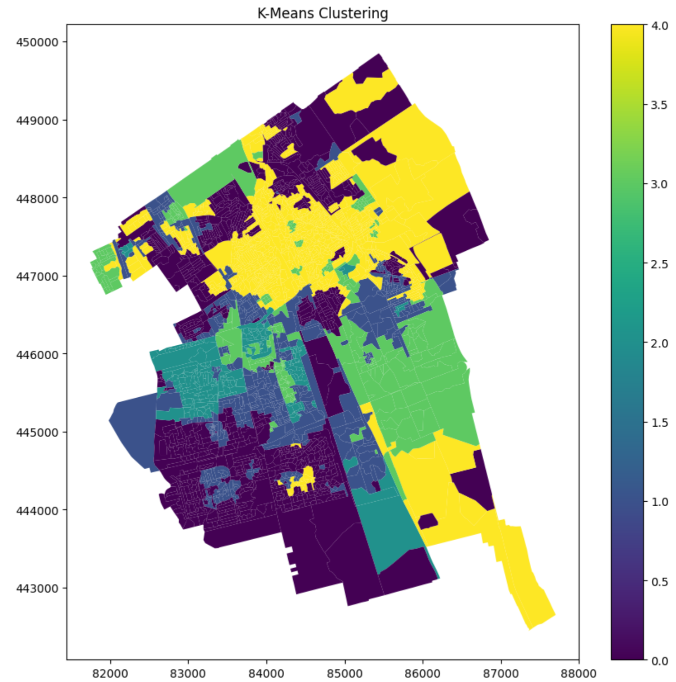


Fig. 9. Result from K Means Clustering - The City of Delft

B. K Means

References

[1] G. Bolt, "Combating residential segregation of ethnic minorities in european cities," *Journal of Housing and the Built Environment*, vol. 24, pp. 397–405, 2009. [Online]. Available: https://doi.org/10.1007/s10901-009-9163-z

[2] L. Spierenburg, S. van Cranenburgh, and O. Cats, "Characterizing residential segregation in cities using intensity, separation, and scale indicators," *Computers, Environment and Urban Systems*, vol. 103, p. 101990, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0198971523000534

[3] T. P. D. Team, *Pandas: Powerful data structures for data analysis, time series, and statistics*, Zenodo, 2020, version latest. [Online]. Available: https://github.com/pandas-dev/pandas

[4] K. Jordahl, J. V. den Bossche, M. Fleischmann, J. Wasserman, J. McBride, J. Gerard, J. Tratner, M. Perry, A. G. Badaracco, C. Farmer, G. A. Hjelle, A. D. Snow, M. Cochran, S. Gillies, L. Culbertson, M. Bartos, N. Eubank, M. Albert, A. Bilogur, S. Rey, C. Ren, D. Arribas-Bel, L. Wasser, L. J. Wolf, M. Journois, J. Wilson, A. Greenhall, C. Holdgraf, and F. Leblanc, *GeoPandas: Python tools for geographic data*, GeoPandas Development Team, 2020, version 0.8.1. [Online]. Available: https://github.com/geopandas/geopandas

[5] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[6] Open Source Geospatial Foundation Project, "QGIS Geographic Information System," http://qgis.org, 2023.

[7] J. Demšar, T. Curk, A. Erjavec, Črt Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Štajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik, and B. Zupan, "Orange: Data mining toolbox in python," *Journal of Machine Learning Research*, vol. 14, pp. 2349–2353, 2013. [Online]. Available: http://jmlr.org/papers/v14/demsar13a.html

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[9] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML*

PKDD Workshop: Languages for Data Mining and Machine Learning, 2013, pp. 108–122.

[10] G. Delft, “Voordijkshoorn,” https://www.delft.nl/voordijkshoorn, accessed on November 6, 2023.

[11] ——, “Kuyperwijk,” https://www.delft.nl/kuyperwijk, accessed on November 6, 2023.

[12] DUWO. (2023) Priority rules for international students. [Online]. Available: https://www.duwo.nl/en/i-search/student-room/are-you-searching-for-a-room/priority/translate-to-engels-voorrangsregeling-voor-internationale-studenten

[13] Room.nl. (2023) Questions and answers. [Online]. Available: https://www.room.nl/en/information-contact/question-and-answer/#/node/1701