Logistic Profile Generation via Clustering Analysis

Andres Regal

***Abstract*—The process of characterizing a city to generate logistic profiles involves the analysis of many different aspects. These profiles are based on secondary sources of data, mainly road network infrastructure, socio-economic data and population density. Following previous research, the final profiles are given by a K-Means algorithm, which uses principal component analysis (PCA) for correlation analysis. A caveat in this method is that prior research has shown that PCA is sensitive to outliers and high dimensionality, which may mislead the following analysis and research. As such, this paper proposes a methodology to evaluate the performance of different clustering techniques to generate logistic profiles, applying it to a case study in the city of Lima, Perú.**

***Index Terms*—Clustering analysis, last mile logistics, logistics, territorial intelligence.**

I. Introduction

The world's urban population has been growing steadily for the last decades [1]. The most recent figures from the UN Department of Economic and Social Affairs set up urban population to grow by 65 million per year, where emerging markets absorb most of the growth, specifically on population density and infrastructure.

As a city grows, the interactions between the stakeholders proposed by [2] become more and more complex. Government officials look to ensure low transportation costs to make the city more competitive and attractive, while residents look to have a high quality of life, which from a logistics perspective may be focused around a greener city with high product availability.

Urban population growth directly translates into an increased demand for goods and services (and the logistic activities that support the supply of these goods). As such, even though most of the stakeholder's interests may be met [2], the influence of urban planning and policy making may compromise the sustainability and "liveability" of a city.

Within this context, the different layers of complexity added to last mile operations have made urban logistics an interesting field that benefits from the input of different disciplines. The competition for scarce resources, such as parking spaces and the road network itself, cause different externalities for the city, mainly global and local pollutants and noise.

Focusing on the business perspective, last mile operations become extremely hard to manage and costly, especially once the routing and warehouse location decisions must take into account urban planning decisions such as time windows, low emission zones and the different tolls and fares applied to freight transport.

This situation, coupled with recent development in analytical solutions for urban areas, make logistic profiles interesting within the scope of Territorial Intelligence (TI). TI focuses on the applicability of said solutions, whether they are destined for practitioners or researchers, to understand the underlying behaviors of the transport system of a city [3].

Recent efforts in developing TI tools focus on Geographic Information Systems (GIS) and collaborative decision making approaches, which are mainly applied to urban land planning and urban transport (including urban logistics) [4]–[6].

Even though the main indicators and software have been thoroughly developed, there is an existing need for data standardization and unification. Analytic approaches to generate logistic profiles, which show a strong potential in supporting decisions regarding urban logistics planning [3], [7], [8], depend strongly on the databases that support them.

To overcome the challenges faced when collecting data for cities in emerging markets, open data sources like LandScan [9] and OpenStreetMaps [10] have been used to produce a high level logistic profile which provides insight into how different zones within the city behave and their similarities [7], [8], [11].

As such, following the efforts of [8], [11], this paper looks to implement a methodology for logistic profile generation based on neural network clustering [12]. The resulting profiles will be compared to the results from both of these works, adding validation indices designed for clustering analysis, to determine the best alternative for the city of Lima.

The remaining sections of this paper are as follows: Section II presents a literature review focused on applications of clustering and zoning approaches in urban logistics, Section III presents the methodological framework for this paper, Section IV discusses the main findings and results and Section V presents the main conclusions of this work.

II. Literature Review

Clustering analysis has mainly been applied to vehicle routing and facility location tasks (within the context of urban freight transport and logistics). Such applications vary between techniques and databases, but the core objective is to cluster a set of clients or target facilities.

As such, Affinity Propagation (AP) [13], is applied by [14] to Vehicle Ad Hoc Networks (VANETs). This is done to address challenges a VANET faces by clustering the network, which in turn will help reduce risk of accidents and traffic congestion. Similarly, [15] applies AP to a fixed-charge facility location problem. The end result allows for facilities to be constructed at specific nodes in a clustered network,

Manuscript received August 23, 2019; revised January 12, 2020.

Andres Regal is with Universidad del Pacifico, Lima, Peru (e-mail: a.regalludowieg@up.edu.pe).

doi: 10.18178/ijmlc.2020.10.1.921

which brings two main benefits: small construction costs and good service to client nodes.

On the use of density based techniques, [16] applies DBSCAN [17] with a particle swarm optimization algorithm to modern logistics and vehicle distribution. This is done by using DBSCAN to cluster routes over a network and applying a Particle Swarm Optimization algorithm to calculate the length of each route, average travel time and to compare the results of this approach with the application of the ant colony algorithm.

Focusing on green vehicle routing, [18] applies DBSCAN in a Green Vehicle Routing Problem (G-VRP), looking to use clustering as a heuristic for large scale routing problems considering fuel restrictions. [19] extends DBSCAN for site location of express enterprises, to cluster together dense regions of costumers for a facility to be placed, which will help reduce transportation costs around the corporate sites. DBSCAN and HDBSCAN [20], the second density based alternative, are compared in [21] when clustering traffic accidents in urban areas. This analysis brings to light some issues with these algorithms: their sensitivity to parameter selection. This problem is addressed by implementing an extension to the algorithms that leads to a more unsupervised parameter selection.

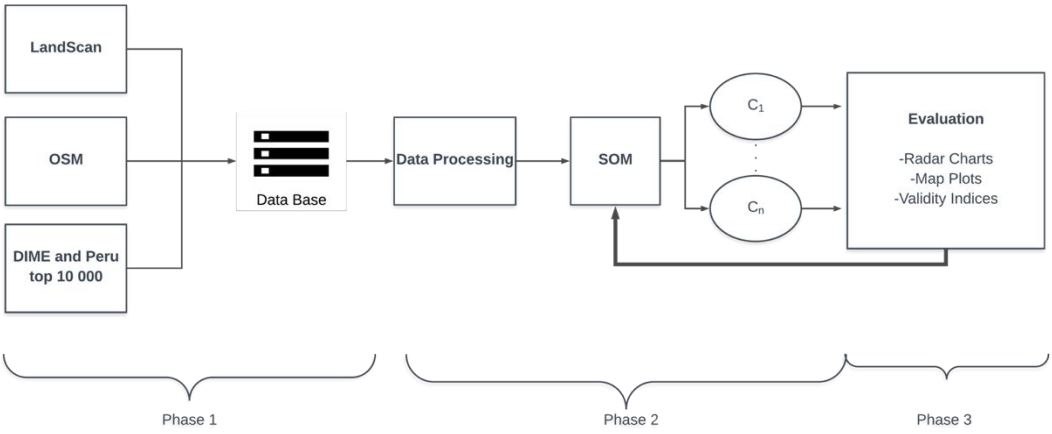

Fig. 1. Proposed methodology.

Finally, zoning approaches for freight trip generation [22], [23] have looked into the use of zoning techniques and data collection frameworks for large urban areas, particularly based on land use and social variables.

Describing the main computational framework applied in this work, two main techniques need to be introduced: Self Organizing Maps (SOM) and Uniform Manifold Approximation and Projection (UMAP). SOM, as defined by [12] is a competitive learning neural network approach to perform clustering analysis. By mapping an n-dimensional input vector to a two-dimensional Kohonen layer, multiple neurons compete to receive an activation signal, which is only given to the neuron which is closes to the input vector (based on a specified distance metric). The success of this technique is based on the lateral interaction of the neurons, enforced by the concept of neighborhood cells [12]. This concept refers to updating the weights of all neurons within a defined neighborhood radius of the "wining" cell, which decreases as training continues.

UMAP, introduced by [24], looks to find a topologically accurate low dimension representation of high dimensional data. This is done via the construction of a weighted graph and minimizing the cross entropy within the high dimensional and low dimensional graphs. This approach looks to preserve the structure of data points in high dimensions by taking local connectivity assumptions within the graph, such that the resulting low dimensional representation preserves the spatial characteristics of the data points.

III. Methodology

The methodology for this paper consists of three stages, visualized in Fig. 1. The first stage, the data collection stage, consists of collecting different secondary sources of data from open access sources. The data collected looks to describe different characteristics from the city from a logistic and urban freight perspective. As such, the dataset has three main components to describe 1 km^2 zones (as per [11]): population density, road network infrastructure and socio-economic variables.

Population levels are processed from LandScan [9] satellite images. These images are provided in raster formats, where each pixel value represents a 1 km^2 area and the intensity value in the image represents the population level. By filtering the image pixels to the limits of Lima, the filtered image is turned into database format, where geographic position and population levels are the main features.

Using the coordinates of these pixels, road network infrastructure data is collected by querying OpenStreetMaps [10] with the OSMnx Python package [25]. These queries construct multiple coefficients, calculated for the road infrastructure contained in the 1 km bounding box of each pixel. Metrics such as the total number of intersections, highway length within the bounding box, the total number of one-way streets and the betweenness centrality and circuity factor coefficients for the bounding box segment are collected. For further detail into the metrics the authors refer the reader to [25].

The third component is the most difficult to collect for a city in emerging markets. Due to the high levels of informality, data regarding the commercial and employment levels is often biased or incomplete. Thus, a combination of Peru's top 100 largest businesses directory and the directory of small and medium businesses (DIME by its acronym in Spanish) is used as an approximation. Each company is

categorized into Food, Beverage and Accommodation, Manufacturing, Retail and Other Services. The total amount of businesses and employees per category is aggregated with respect to the 1 km^2 areas. Finally, the dataset for the next stage consists of 652 records with 17 features.

The second stage consists of data processing and clustering. As mentioned before, this paper will focus on a neural network approach to cluster the city zones. Specifically, Self-Organizing Maps (SOM) [12] will be the main algorithm to be tested. Leveraging the results from [8], [11], the components produced using PCA [26] and UMAP [24] will be used as inputs as well as the raw features to test how different dimension reduction approaches influence the end result.

The final stage consists of an evaluation of the validity and coherence of the results. To perform this evaluation, the Calinski-Harabasz [27] and the Silhouette Coefficient [28] will be calculated for each experiment, as well as a geographical evaluation of the coherence of the profile.

IV. Results

In this section the main results of this paper are presented following the application of aforementioned methodology. Analyzing the logistic profile of Lima requires an initial understanding of its relevance to Peru. As a megacity with 10 million inhabitants, Lima concentrates approximately 30% of Peru's population and it holds both the largest airport and port. Administratively, the city is divided into 49 districts (43 in Lima and 6 in Callao) which act independently.

To visualize the segmentations proposed by each experiment, there are two main alternatives. First, as in Fig. 2, a geographical plot may be used to visualize the location of each 1 km^2 zone and its respective cluster (as a color). Following this approach, it is evident that the choice of using a dimension reduction technique, and the decision of which reduction method to use, has great impact in the end results of the clustering analysis.

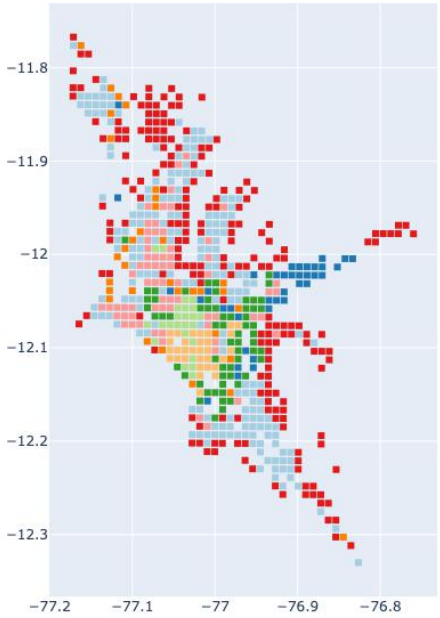

Fig. 2. Geographical plot for PCA clustering.

All approaches converged to 8 different zones. Using PCA in conjunction with SOM, the zones have strong functional relationships (since features regarding geographical coherence have low variance). As shown in Fig. 2, the outskirts are grouped into a single cluster, then the northern and southern poles of the city are roughly aggregated into the same cluster, followed up by small divisions within the core of the city.

Given the prevalence of the functional features, spatial coherence is not a priority for the SOM, especially as the analysis shifts away from the historic center and the financial district (light green and orange, respectively). Even so, PCA is able to capture the most complex behaviors in the city. As shown in Fig. 3, cluster 6 (a combination of the financial, commercial and historic centers) captures the highest level of commercial establishments with a reasonably high level of network complexity.

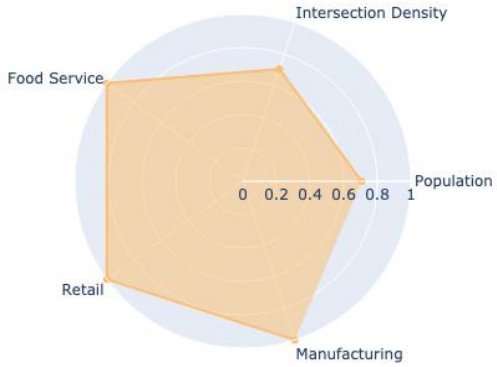

Fig. 3. Main variable plot for Cluster 6.

It is also important to note, as shown in Table I, that using PCA (looking for 95% variance) results in the worst performing scores of the three experiments. Analyzing the silhouette score for this approach, the 0.2731 silhouette score and the 213.80 Calinski-Harabasz score conveys a configuration in which data points have high intra-cluster variance and low inter-cluster variance.

The second, and best performing, experiment was UMAP in conjunction with SOM. In this case, the low dimensional embedding of the characteristics of each zone results in spatially contiguous zones which maintain functional relationships.

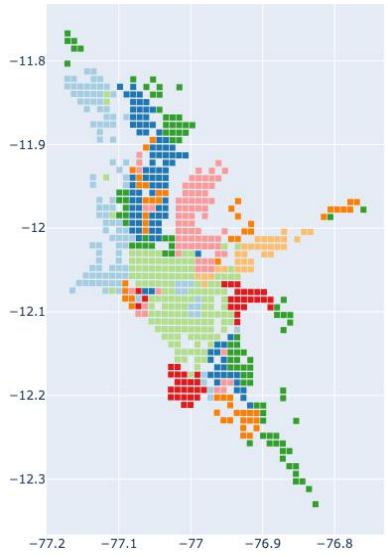

Fig. 4. Geographical plot for SOM clustering with UMAP components.

Geographically, this spatial coherence results in two districts getting clustered on their own. One of those, San Juan de Lurigancho (pink in Fig. 4), is the most densely populated district in Lima. Given the high population levels, strong presence of commercial establishments and the spatial characteristics, separating these districts as its own cluster reflects a real behavior of the city. This is the same case as Callao, which gets clustered on its own. Callao's case also represents an administrative segmentation within Lima, which may also be interpreted as an adequate segmentation.

Another interesting result from this algorithm comes from the second cluster, shown in Fig. 5. Commercially it is very similar to what the PCA approach found, but the spatial component causes the financial, commercial and historic centers to merge into a single cluster, including some residential areas (increasing the intersection and population densities).

TABLE I: CLUSTERING VALIDATION INDICES

Metric	SOM with PCA	SOM with UMAP	SOM with raw features
Calinski-Harabasz	213.80	**5719.03**	3017.98
Silhouette Score	0.2731	**0.7737**	0.6159

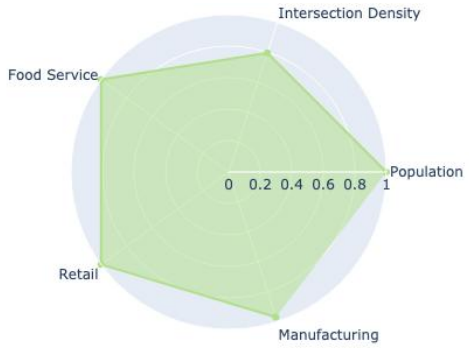

Fig. 5. Main variable plot for Cluster 2.

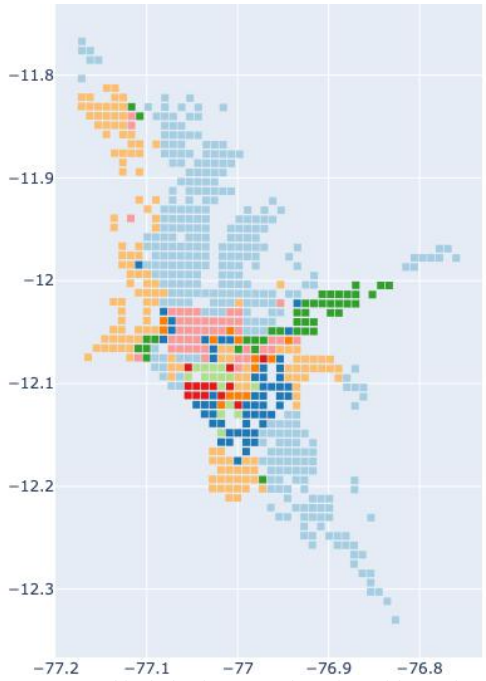

Fig. 6. Geographical plot for SOM clustering with raw features.

The third and final experiment consists of using the raw features prior to the dimension reduction. Within this high dimensional approach and given that SOM does not use local connectivity assumptions, small differences and correlations between the input features cause large districts to be clustered together, instead of showing more micro behaviors for the 1 km^2 zones.

The large coupling of 1 km^2 zones following administrative districts allows higher clustering scores, but the small differences caused by clusters 4, 5 and 2 capture the micro behaviors which are interesting from an urban logistics perspective (Fig. 6). Thus, as shown in Fig. 7, cluster 5 captures the same commercial behavior while grouping the residential zones into cluster 2.

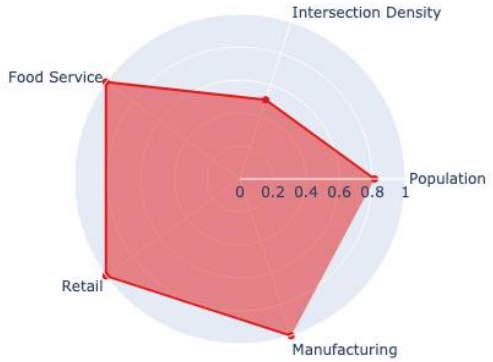

Fig. 7. Main variable plot for Cluster 5.

Being able to capture the micro behavior and grouping large areas adds a layer of complexity to the analysis of the profile. Even though grouping large sets of zones with similar characteristics allows for targeted policy making to focus on the most commercial and active zones of the city, within the large district clusters the data collection and validation may not be the best.

When dealing with high dimensional data, missing values or outliers (usually to the lower ends of the distribution of each variable) from the open data sources may cause these large clusters. Also, these large couplings into administrative regions may point to weak data collection efforts within these areas, highlighting the need for city-wide data collection and validation efforts to have a more realistic view of the system's behavior.

Finally, the unified distance matrices (U-matrix) can bring further insight into the inner workings of the SOM results. Using the u-matrix and plotting the position of each data point in this 2D map, the clusters can be identified based on the activations of the neurons (higher color intensity). In the figures below, lower activation values correspond to lighter colors (following the spectral color map) and darker colors to higher activation.

As a general interpretation rule, lower activations represent a large distance between neurons and thus a gap between the codebook values in the input space. A dark coloring in the plane signifies that the codebook vectors are close to each other in the input space. As such, dark areas can be thought as clusters and light areas as cluster separators.

The u-matrix for the PCA based SOM is presented in Fig. 8, where the different data points are distributed uniformly, with no prominent activations to split the clusters (visually, there are no distinct groups when projecting the data points in the u-matrix surface). This is a first indicator of why the validation scores tend to be lower than in other approaches,

clusters are less separated within this 2D grid and the absence of a linear correlation within the input features may make learning this plane more complex.

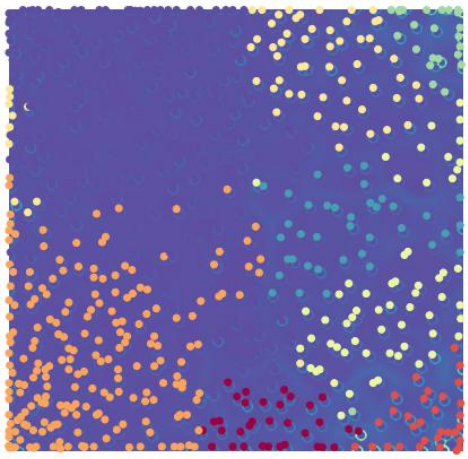

Fig. 8. U-Matrix for SOM using PCA components.

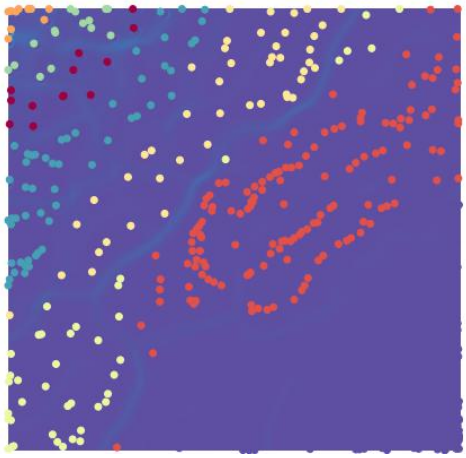

Fig. 9. U-Matrix for SOM using raw features.

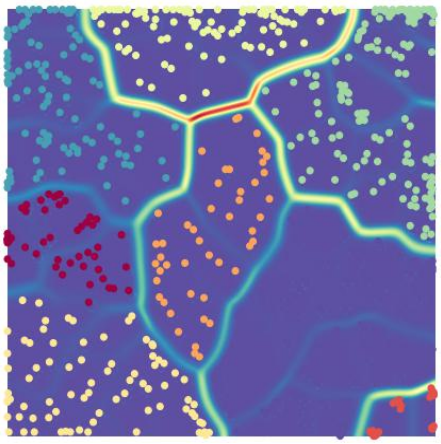

Fig. 10. U-Matrix for SOM using UMAP components.

When plotting the u-matrix for the raw feature experiment, as shown in Fig. 9, SOM is capable of finding a representation which splits the data with diagonal traces along the plane. There are light activations which serve as separators, but the projection does not relate to well-formed groups aside from those at the far right of the figure.

The third approach, as shown in Fig. 10, does show a structured plane and neuron activations splitting the clusters. The embedding procedure performed by UMAP allows SOM to learn a clear 2D plane in which the activations split the data points clearly, since the high dimensional information from the original data set is maintained through the low dimensional graph representation. This results in clusters with low intra cluster variance and high inter cluster variance, hence the highest validation scores and spatial coherence.

V. Conclusions

The conclusions of this paper may be outlined as follows. A competitive learning approach, following previous research con the topic of logistic profile generation, has shown that the use of dimension reduction techniques such as UMAP can increase the performance of algorithms like SOM thanks to the low dimension embedding of original high dimensional feature space.

Another implication relays on this paper's results guiding the deployment of territorial analytics indicators and urban freight models specifically designed with zoning approaches. Given the spatially and functionally correlated clusters, specific land use, socio-economic or environmental indicators may be deployed to assess the performance of certain city areas.

A final conclusion is that the proposed zoning, with focus con spatial or functional characteristics, may be used as a decision support system when designing public policy regarding urban logistics. Since different regions of a city behave in similar ways, data driven decisions are key to guaranteeing a sustainable and livable city.

Author Contributions

A. Regal worked on data collection, processing and writing the paper.; all authors had approved the final version.

References

[1] United Nations, "Concise Report on the World Population Situation in 2014," *Dep. Econ. Soc. Aff. Popul. Div.*, pp. 1–38, 2014.

[2] E. Taniguchi, "City logistics for sustainable and liveable cities," *Green Logist. Transp. A Sustain. Supply Chain Perspect.*, vol. 151, pp. 49–60, 2015.

[3] J. González-Feliu, "Urban logistics and spatial territorial intelligence indicators: State of the art, typology and implications for Latin American cities," *Interfases*, no. 001, pp. 135–176, 2018.

[4] T. T. Nguyen, P. Krishnakumari, S. C. Calvert, H. L. Vu, and H. V. Lint, "Feature extraction and clustering analysis of highway congestion," *Transp. Res. Part C Emerg. Technol.*, vol. 100, pp. 238–258, 2019.

[5] J. Zhao, H. Xu, H. Liu, J. Wu, Y. Zheng, and D. Wu, "Detection and tracking of pedestrians and vehicles using roadside LiDAR sensors," *Transp. Res. part C Emerg. Technol.*, vol. 100, pp. 68–87, 2019.

[6] J. Fan, C. Fu, K. Stewart, and L. Zhang, "Using big GPS trajectory data analytics for vehicle miles traveled estimation," *Transp. Res. Part C Emerg. Technol.*, vol. 103, pp. 298–307, 2019.

[7] D. Merchán, *El Perfil Logístico de Quito*, pp. 1–50, 2015.

[8] A. Regal, J. Gonzalez-Feliu, M. Rodriguez, and M. Mathieu-Jugunaru, "Defining urban logistics profile zones in South American metropolis by combining functional and spatial clustering techniques," in *Proc. the 3rd International Conference on Control, Automation and Diagnosis*, pp. 161–174.

[9] A. N. Rose and E. A. Bright. (2014). The LandScan Global Population Distribution Project: current state of the art and prospective innovation. [Online]. Available: https://pdfs.semanticscholar.org/dbec/08b982769c197b8b891390e55e055581c5db.pdf

[10] OpenStreetMap contributors. (2017). Planet dump. [Online]. Available: https://planet.osm.org

[11] M. Winkenbach *et al.*, "City logistics policy toolkit: A study of three latin american cities," Final Report, 2018.

[12] T. Kohonen, "Exploration of very large databases by self-organizing maps," in *Proc. International Conference on Neural Networks (ICNN'97)*, 1997, vol. 1, pp. PL1--PL6.
[13] D. Dueck and B. J. Frey, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.
[14] C. Shea, B. Hassanabadi, and S. Valaee, "Mobility-based clustering in VANETs using affinity propagation," in *Proc. Globecom - IEEE Glob. Telecommun. Conf.*, 2009.
[15] W. Li, L. Xu, and D. Schuurmans, "Facility locations revisited: An efficient belief propagation approach," in *Proc. 2010 IEEE Int. Conf. Autom. Logist. ICAL 2010*, 2010, pp. 408–413.
[16] H. Shi, Z. Li, W. Li, and Z. Yu, "Application of particle swarm optimization based on clustering analysis in logistics distribution," in *Proc. 2009 Int. Conf. Manag. e-Commerce e-Government*, 2009, no. 1, pp. 291–295.
[17] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, vol. 54, 2nd ed., 2006.
[18] S. Erdoğan and E. Miller-Hooks, "A green vehicle routing problem," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 48, no. 1, pp. 100–114, 2012.
[19] W. P. Wang and B. Yang, "Site location of express enterprise through DBSCAN-based spherical clustering," in *Proc. 2008 Int. Conf. Wirel. Commun. Netw. Mob. Comput. WiCOM 2008*, pp. 1–5, 2008.
[20] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," *Adv. Knowl. Discov. Data Min.*, pp. 160–172, 2013.
[21] E. Rosalina, F. D. Salim, and T. Sellis, "Automated density-based clustering of spatial urban data for interactive data exploration," in *Proc. the IEEE International Conference on Computer Communications (INFOCOM) Workshops*, May 2017.
[22] C. T. Lawson, E. L. Powers *et al.*, "Estimated generation of freight trips based on land use," *Transp. Res. Rec.*, vol. 2269, no. 1, pp. 65–72, 2012.
[23] J. Holguin-Veras and M. Jaller, "Comprehensive freight demand data collection framework for large urban areas," *Sustainable Urban Logistics: Concepts, Methods and Information Systems*, Springer, 2014, pp. 91–112.
[24] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv Prepr. arXiv1802.03426*, 2018.
[25] G. Boeing, "OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks," *Comput. Environ. Urban Syst.*, vol. 65, pp. 126–139, 2017.
[26] I. Jolliffe, *Principal Component Analysis*, Springer, 2011.
[27] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Stat.*, vol. 3, no. 1, pp. 1–27, 1974.
[28] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. C, pp. 53–65, 1987.

Andres Regal is a research assistant at Universidad del Pacífico in Lima, Perú. He has worked as a data scientist at Corporación Breca, Perú and got his BSc in data science from Universidad del Pacífico. His main research interests focus on smart cities and intelligent transportation systems, focusing on analytical zoning approaches and data-driven policy making.