Spatial and aspatial clustering analysis of PM2.5 concentrations in Temuco, Chile using mobile measurements

Carola A. Blazquez, PhD and Elizabeth Montero, PhD
Department of Engineering Sciences, Universidad Andres Bello, Viña del Mar, Chile
{cblazquez, elizabeth.montero}@unab.cl

***Abstract*—Air pollution due to wood burning produces severe health and environmental problems. Clustering methods are needed to estimate PM2.5 exposures, and identify locations with high PM2.5 concentrations. This study performed a spatial and aspatial clustering analysis of PM2.5 pollutant collected in a mobile campaign in the conurbation of Temuco and Padre Las Casas, Chile. The Getis Ord G_i^* statistic was employed to obtain spatial variability of PM2.5 concentrations, and a K-Means clustering method was used to group PM2.5 concentrations with a aspatial perspective. In addition, an integrated spatial and aspatial clustering approach was implemented with the PM2.5 concentration and measurement spatial location. The comparison results suggest that integrating the spatial and aspatial clustering methods yield high quality partitions when considering spatial information.**

***Index Terms*—Clustering, PM2.5 concentrations, mobile measurements, K-Means, Getis-Ord G_i^***

I. Introduction

Ambient air pollution causes serious human health problems such as cardiovascular and chronic respiratory diseases and lung cancer, as well as adverse environmental effects [1]–[3]. According to the World Health Organization, every year an estimated 4.2 million deaths worldwide are attributable to ambient air pollution exposure, particularly due to PM2.5, particulate matter less than 2.5 micrometers in diameter [4], [5]. PM2.5 is capable of reaching the respiratory track and producing damages in different parts of the body through air exchange in the lungs [3]. Epidemiological studies have revealed that PM2.5 increases the risk of morbidity and even premature mortality [1], [6]. PM2.5 is emitted into the air as fine particles from residential wood combustion generated from cook stoves and space heaters. This wood combustion occurs typically in urban areas and is a relatively cheap fuel compared to diesel and gasoline or oil combustion [7], [8]. Wood burning pollution is a critical problem in southern Chile, in which approximately 2 million inhabitants are exposed to high environmental risk [9]. In particular, Temuco and the adjoining commune of Padre Las Casas (PLC) present a severe wood burning pollution problem with average daily levels of PM2.5 concentrations of approximately 430 ug/m3. Note that the Chilean normative declares an environmental emergency when the level of PM2.5 concentrations in 24 hours is equal to or greater than 170 ug/m3 [10]. Over 80% of the PM2.5 emissions are caused by residential wood burning mainly in the winter season and at night when residential heating increases and atmospheric dispersion decreases [8], [9], [11], [12]. These high PM2.5 emissions are due to inefficient wood stoves, poor household insulation, and lack of adequate wood burning practices [8]. Therefore, there is a need for identifying the spatial variability of different PM2.5 concentration levels in the conurbation of Temuco and PLC.

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2020.1.1.528
ISBN: 978-958-52071-4-1 ISSN: 2414-6390

II. Literature review

Different methodologies have been employed to estimate PM2.5 exposures at different locations for epidemiological and environmental studies, and thus, capture the spatial variability of PM2.5 concentrations [7]. For example, some studies have used Kriging as a geostatistical technique [13]–[15], Land-Use Regression [16]–[18], and clustering techniques [19]–[21] to model spatial variation in air pollution concentrations.

In particular, clustering methods are widely used to recognize homogeneous groupings of data [22]. These methods may be classified into aspatial and spatial clustering. While aspatial clustering approaches utilize solely independent point values when partitioning the data into clusters, spatial clustering methods consider the degree of similarity among neighboring features (i.e., spatial autocorrelation) to determine the spatial dependence or independence among the features [22].

Studies have combined spatial and aspatial clustering techniques in different fields of study, proving overall improved results. For example, [22] concluded that integrating K-Means clustering and the spatial Getis-Ord G_i^* statistic resulted in a superior technique for identifying clusters of orchards. In another study, [23] employed multiple K-Means clustering and Moran's I spatial autocorrelation method to find the preferences and provincial characteristics of consumers in the retail industry. In [24] Scrucca implemented K-Means clustering and measures of spatial autocorrelation to incorporate the spatial structure of the labor market data in Umbria, Italy. In order to determine irrigation management zones. In [25] Ohana-Levi et. al. employed a weighted multivariate spatial clustering model that integrates Getis-Ord G_i^* statistic and K-Means clustering. Finally, [26] analyzed satellite images to identify local aridity in Indonesia based on vegetation indices by combining spatial autocorrelation (Moran's I statistic) and K-Means clustering.

Few studies have been found in the literature to combine spatial and aspatial clustering methods to identify clusters of PM2.5 concentrations. For instance, [27] employed a local spatial autocorrelation model and hierarchical clustering analysis to classify the monthly average PM2.5 concentrations, to determine the spatial distribution of PM2.5, and thus, to identify the cities with the highest PM2.5 concentration.

In this study, we performed a local spatial autocorrelation using Getis-Ord G_i^* index to identify spatial clusters of PM2.5 concentrations. Subsequently, K-Means technique was employed for the aspatial clustering of this pollutant, and an integrated spatial and aspatial clustering method (both Getis-Ord G_i^* and K-Means) was applied to assess the extent of spatial variability of PM2.5 concentrations in the conurbation of Temuco and PLC due to woodsmoke generation. A comparison analysis was performed to examine the clustering quality among the clustering methods.

III. Data

A. Study area

Temuco is the main city of the region of Araucanía in southern Chile with a population of 282,415 inhabitants [28]. Approximately 23% of the households are classified as poor with an average of 8.2 schooling years of the head of the household [9]. The surrounding area is dominated by forest and a touristic lake district. The region of Araucanía is a major producer of crops, fruits, cattle ranching, and forestry [29]. Although the industrial activity is relatively low, an important air pollution source includes industrial boilers that use wood and coal as fuel. However, residential wood burning is the largest source of PM2.5 in the study area since approximately 90% of the households in Temuco and PLC have woodstoves [9].

In this study, a combined spatial and aspatial clustering analysis and comparison are performed for the whole study area, and also separately for each of the five collection zones depicted in Fig. 1.

B. Data collection and description

Mobile sampling has been used to describe and characterize the spatial distribution of air pollutants [30], [31]. The advantage of mobile monitoring campaigns of air pollutants is that achieves unparallel spatial coverage when compared to fixed-site sampling, and it is a simple and economical manner for spatial distribution exploration [11], [30]. In this study, a mobile sampling campaign was performed in Temuco and nearby commune of PLC during 20 nights in the winter of 2016 between 8:00pm and midnight following different routes to cover each of the five collection zones shown in Fig. 1.

The instrumentation consisted of two DustTrack II to collect mobile and fixed PM2.5 concentrations and a GPS receiver to obtain position coordinates every second. Note that a calibration phase took place using a fixed central site prior to initiating the mobile measurements. The PM2.5 measurements were normalized with the background concentrations at the fixed central site to minimize the influence of the meteorological variations in the air pollution [31].

Table I presents basic descriptive statistics of the collected PM2.5 concentrations for the whole city and each collection zone. This table indicates that on average the highest and lowest PM2.5 concentrations were measured in Las Encinas and PLC, respectively.

TABLE I: Data summary

Area	PM2.5 Concentrations				
	# Measurements	Mean	S.D.	Min	Max
Whole study area	162.4	127.1	105.3	2.6	1,255.3
Collection zones					
Amanecer	40.3	104.4	66.6	2.6	933.7
Labranza	13.8	110.6	90.6	13.4	704.3
Las Encinas	56.9	173.4	140.2	6.7	1,255.3
Padre Las Casas	18.9	85.1	55.1	3.9	431.9
Pedro de Valdivia	32.4	105.8	65.9	4.9	517.1

IV. Methodology

A. Local spatial autocorrelation

Local spatial autocorrelation identifies statistically significant spatial clusters with high (hotspots) and low (coldspots) values. The Getis-Ord G_i^* statistic is commonly used to identify these hotspots and coldspots by testing the null hypothesis that the spatial autocorrelation of a variable is equal to zero. If the null hypothesis is rejected, then the variable is spatially autocorrelated (Ord and Getis, 1995). The Getis-Ord G_i^* statistic is expressed by (1).

$$G_i(d) = \frac{\sum_{j=1}^{n} w_{ij}(d)x_j - \bar{x}\sum_{j=1}^{n} w_{ij}(d)}{s\sqrt{\frac{n\sum_{j=1}^{n} w_{ij}^2 - (\sum_{j=1}^{n} w_{ij})^2}{n-1}}} \quad (1)$$

Where x_j indicates the attribute value at location j, $w_{ij}(d)$ is a spatial weight matrix for all locations j within distance d from the feature at location i, n is the total number of locations, $\bar{x}$ is the sample mean, and s is sample variance.

Z-score and p-values are outputs of the Getis-Ord G_i^* statistic, which indicate measures of statistical significance [32]. The larger (smaller) the Z-score is, the more intense the clustering of high (low) values. Random distribution exists for Z-score values near zero. In this study, Z-score values are classified into three clusters: 0 (negative spatial autocorrelation with Z-score<-1.65 and p-value<0.1), 1 (no spatial autocorrelation with Z-score $\in$ [-1.65, 1.65] and p-value$\geq$0.1), and 2 (positive spatial autocorrelation with Z-score>1.65 and p-value<0.1).

B. K-Means clustering

K-Means is an iterative algorithm that separates a set of data in a set of k groups. Starting from a set of k centroids, K-Means minimizes the within-cluster sum of squares. Similarity between observations is measured as Euclidean distance between attribute values [33]. At each step, centroids are updated as the mean of the observations in each cluster. As new observations are included in clusters, centroids are updated.

K-Means is one of the most used clustering method. It is simple to implement and has a lineal complexity, but has

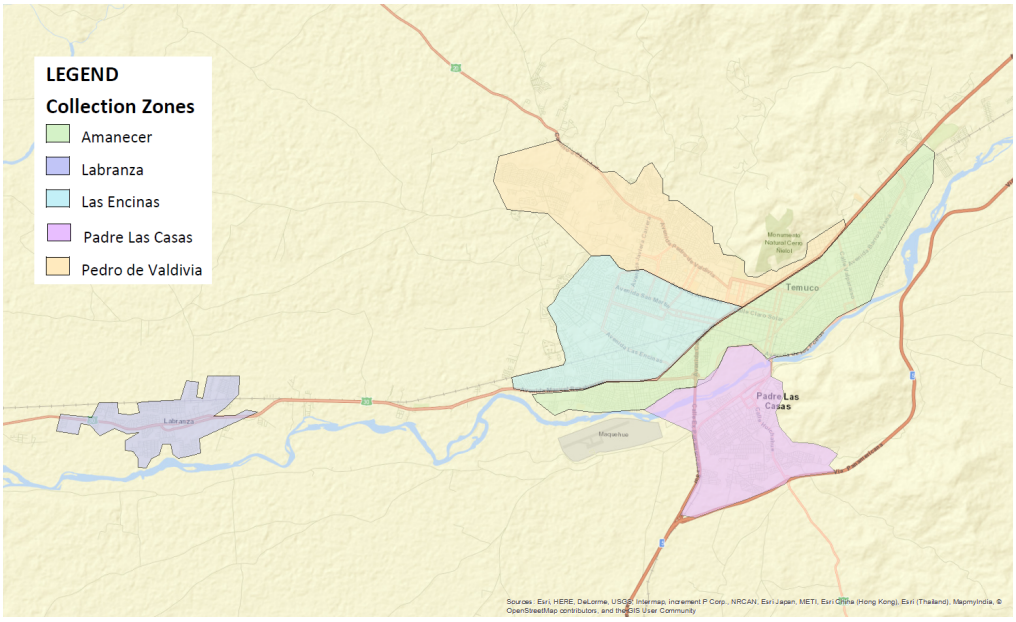


Fig. 1: Collection zones of PM2.5 concentrations

some disadvantages (e.g., no convergence can be assured, the number of clusters is a user defined parameter of the method, and is sensitive to the presence of outliers in the data).

V. Experimental Results

This section presents the results of the spatial and aspatial clustering, and also the integrated spatial and aspatial approach that were applied to the PM2.5 concentrations measured in Temuco and PLC as a whole study area, and also in each of the five collection zones.

In the local spatial autocorrelation analysis, the zone of indifference method was employed to determine the spatial relationship among the PM2.5 measurements. In this method, a influence weight equal to one is assigned to those measurements within the threshold distance d of the measurement under study. This weight decreases with distance for those measurements that are located beyond the threshold distance. The threshold distances employed for each zone are shown in Table II.

TABLE II: Threshold distances for the whole study area and each collection zone

Area	Threshold distance (meters)
Whole study area	27.4
Collection zones	
Amanecer	26.6
Labranza	18.0
Las Encinas	27.4
Padre Las Casas	20.8
Pedro de Valdivia	23.8

The results of the Elbow Method for the K-means clustering analysis indicate that three clusters (k=3) yielded the best results in most cases. Thus, this number of clusters was used in the local spatial autocorrelation and integrated spatial and spatial approach for comparison purposes.

Tables III, IV, and V present the mean and standard deviation of the PM2.5 concentrations for the clusters with low (0), medium (1), and high (2) values that were obtained with the Getis-Ord G_i^* statistic, K-Means, and the combination of both, respectively. Note that, in Table III, low values (0) are coldspots and high values (2) are hotspots of PM2.5 concentrations, while medium values (1) present a random pattern or no clustering.

The aforementioned tables show that the Las Encinas zone has the highest average PM2.5 concentration (even higher than when the whole study area is analyzed), while the PLC zone presents the lowest average values of the PM2.5 pollutant. Las Encinas is located in the center of the city with the lowest elevation and a large upper class population, and PLC is a low class commune that adjoins Temuco in the south at a higher elevation than Las Encinas. This is partly due to the tendency of PM2.5 concentrations to decrease with increasing elevations [34].

TABLE III: Clustering summary – Getis-Ord statistic

Area	0: low		1: medium		2: high	
	Mean	S.D.	Mean	S.D.	Mean	S.D
Whole study area	76.9	47.8	122.8	55.6	204.2	143.3
Collection zones						
Amanecer	64.5	45.8	107.2	58.0	155.7	59.5
Labranza	55.1	28.4	105.7	53.8	198.2	132.4
Las Encinas	111.9	53.5	159.4	82.1	285.6	192.9
Padre Las Casas	44.4	26.5	81.7	41.7	120.3	56.4
Pedro de Valdivia	62.5	36.4	103.9	47.1	153.4	67.5

Figures 2 and 3 present the clustering results of PM2.5 concentrations for the whole conurbation of Temuco and PLC

TABLE IV: Clustering summary – K-Means

Area	0: low		1: medium		2: high	
	Mean	S.D.	Mean	S.D.	Mean	S.D
Whole study area	63.6	29.1	168.3	43.2	494.6	151.3
Collection zones						
Amanecer	45.2	24	136.4	23.2	218.3	40
Labranza	61.8	27.9	171.4	45.6	453.6	106.9
Las Encinas	120.4	46	349.2	79.6	679	140.2
Padre Las Casas	41.7	15.4	95.9	18.2	176.4	37.1
Pedro de Valdivia	59.7	25.6	157.4	29.6	328.1	65.8

TABLE V: Clustering summary – Integrated approach

Area	0: low		1: medium		2: high	
	Mean	S.D.	Mean	S.D.	Mean	S.D
Whole study area	81.3	49.6	146	71.8	308.4	201.1
Collection zones						
Amanecer	41.4	40	79	52.8	152.2	59.3
Labranza	68.4	37.4	135.8	73.9	379.5	171.9
Las Encinas	97.5	36.5	125.2	64.2	273	188.1
Padre Las Casas	36.8	22.2	76.1	40.7	124.3	56.7
Pedro de Valdivia	48.5	19.1	75.6	44.1	149.2	66.5

and separately for each collection zone, respectively. These figures present the low, medium, and high values of PM2.5 concentrations denoted by 0, 1, and 2, respectively. Differences are observed among the clustering methods in both figures. For example, when taking into account the whole study area in Figure 2, a larger zone of high PM2.5 concentration is detected in Las Encinas and part of Pedro de Valdivia (north) and Amanecer (south) zones with the local spatial autocorrelation (Getis-Ord G_i^* statistic). Moreover, Figure 3 shows that high PM2.5 concentrations tend to cluster more when the analysis is performed separately by collection zone than for the whole area. When each collection zone is analyzed separately, the three types of clustering methods group PM2.5 concentrations with less number of measurements, and clusters with low, medium, and high values are identified in a smaller area.

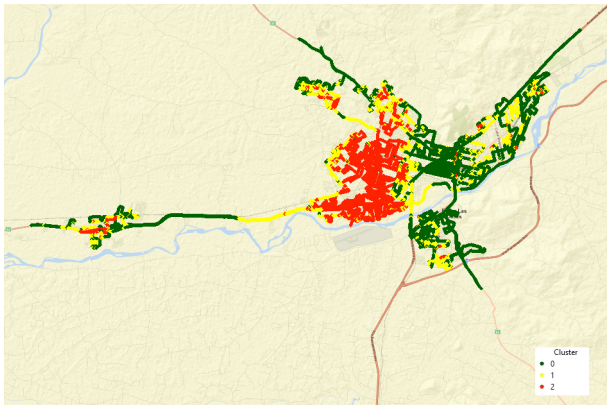
VI. Comparison Results

In order to compare the quality of different clustering processes, we used the Silhouette coefficient [35]. The Silhouette coefficient definition is shown in (2), which measures for each point the relative difference between the average distance to each point within the same cluster and each point in the closest cluster.

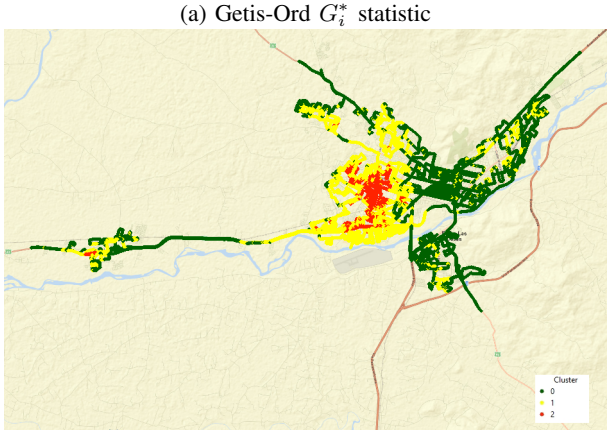
$$S(C) = \frac{1}{n} \sum_{c_k \in C} \sum_{x_i \in c_k} \frac{b(x_i, c_k) - a(x_i, c_k)}{max\{a(x_i, c_k), b(x_i, c_k)\}} \quad (2)$$

where the dataset X is a set of n points and each one is represented as a vector in a F-dimensional space, and k is the number of clusters in the partition. Moreover, equations (3) and (4) show the computations of functions a and b respectively.

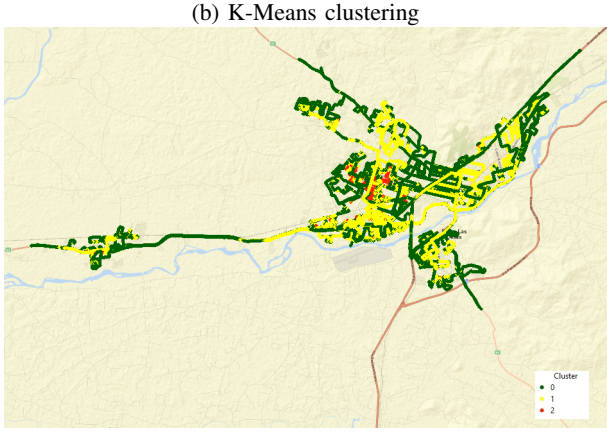
$$a(x_i, c_k) = \frac{1}{|c_k|} \sum_{x_j \in c_k} d_e(x_i, x_j) \quad (3)$$



(a) Getis-Ord G_i^* statistic



(b) K-Means clustering

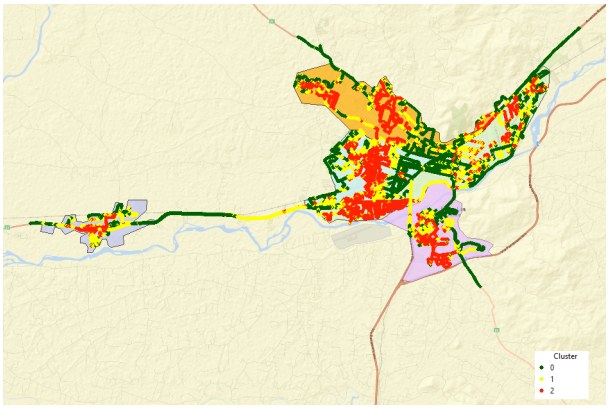


(c) Integrated K-Means and Getis-Ord G_i^*

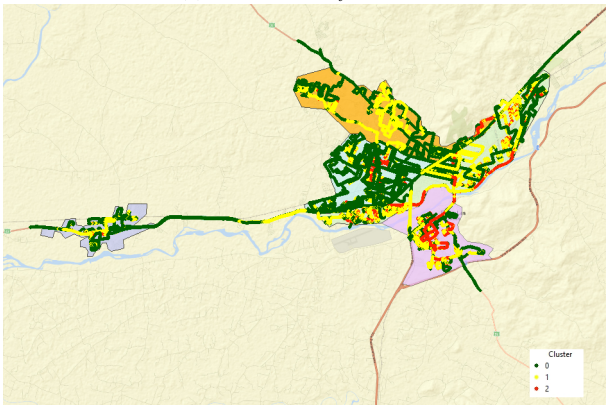
Fig. 2: Clustering for the whole study area

$$b(x_i, c_k) = min_{cl \in C \setminus c_k} \left\{ \frac{1}{|c_l|} \sum_{x_j \in c_l} d_e(x_i, x_j) \right\} \quad (4)$$

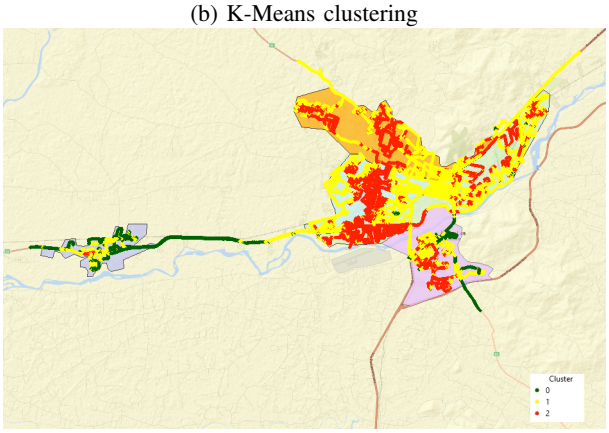
For each point i, the Silhouette coefficient represents how well it was assigned. If $S(i)$ is close to 0, then it is situated at the inflection point between two clusters. If $S(i)$ is close to -1, then we would have been better off assigning the point to the other cluster. If $S(i)$ is close to 1, the point is correctly assigned and can be interpreted as belonging to the appropriate cluster.



(a) Getis-Ord G_i^* statistic



(b) K-Means clustering



(c) Integrated K-Means and Getis-Ord G_i^*

Fig. 3: Clustering for each collection zone

Table VI shows the Silhouette coefficient values for all the cluster partitions that were analyzed in this study. Moreover, for each case, the Silhouette coefficient was computed considering only the PM2.5 concentrations, and also considering the PM2.5 concentrations and the spatial locations of the measurement points (P+L). Data was normalized to avoid bias related to differences in ranges of values. The results in Table VI show that higher Silhouette values were obtained when using only the PM2.5 concentration than when including the location information since all these clustering processes were based on the PM2.5 concentrations. Moreover, according to these values, the best clustering processes were performed by K-Means when using only the PM2.5 concentrations. From the results presented in Section V, we observe that these partitions were unable to clearly differentiate geographical areas with different air pollution levels. From Table VI, when using P+L, we observe that the partitions performed by Getis-Ord G_i^* indicator were the best in three out of six cases and the partitions performed by the integrated approach obtained the best Silhouette indicators in the remaining three cases.

TABLE VI: Silhouette coefficient results

Area	Getis-Ord G_i^*		K-means		Integrated	
	PM2.5	P+L	PM2.5	P+L	PM2.5	P+L
Whole study area	0.186	0.110	0.702	-0.034	0.378	0.016
Collection zones						
Amanecer	0.181	0.090	0.659	0.071	0.121	0.111
Labranza	0.274	0.026	0.729	-0.143	0.431	-0.118
Las Encinas	0.211	0.136	0.755	0.027	0.184	0.024
Padre Las Casas	0.149	0.119	0.669	0.002	0.138	0.154
Pedro de Valdivia	0.216	0.063	0.750	-0.010	0.100	0.127

VII. Conclusions

This study implemented a spatial and aspatial clustering method of PM2.5 concentrations that were collected in a mobile campaign in the conurbation of Temuco and Padre Las Casas in Chile during the winter of 2016. We used Getis-Ord G_i^* statistic to perform the spatial autocorrelation, and thus, hotspots and coldspots were identified throughout the city. K-Means was employed to group the PM2.5 concentrations in three clusters with a aspatial perspective. Subsequently, both Getis-Ord and K-Means were combined to perform the clustering of this pollutant. The Silhouette coefficient was used to compare the results of the quality in the clustering processes. Observing the results on the whole study area, we observe that the use of spatial approaches allow the identification of separated areas of the same hazardous level. This can be noticed when evaluating the Silhouette coefficient values, where the best P+L values were obtained using the spatial approach (higher than zero in both cases). Moreover, analyzing the results from separated collection zones, we observe that the higher P+L Silhouette coefficient values were obtained by the integrated approach in Amanecer, Padre Las Casas, and Pedro de Valdivia.

Future research should include elevation information captured in each GPS measurement to relate the topography of the conurbation with the air pollution. Additionally, other clustering techniques such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) or Agglomerative Hierarchical Clustering will be employed to compare with the current results.

Acknowledgments

Authors want to thank to Dr. Pablo Ruiz from Public Health School at Universidad de Chile and Dr. María Elisa Quinteros from Universidad de Talca for providing the mobile PM2.5 measurement data.

REFERENCES

[1] A. Colao, G. Muscogiuri, and P. Piscitelli, "Environment and health: Not only cancer," *International Journal of Environmental Research and Public Health*, vol. 13, no. 7, 2016. [Online]. Available: https://www.mdpi.com/1660-4601/13/7/724

[2] A. Ghorani-Azam, B. Riahi-Zanjani, and M. Balali-Mood, "Effects of air pollution on human health and practical measures for prevention in Iran," *Journal of Research in Medical Sciences*, vol. 21, no. 1, p. 65, 2016.

[3] Y.-F. Xing, Y.-H. Xu, M.-H. Shi, and Y.-X. Lian, "The impact of pm2.5 on the human respiratory system," *Journal of Thoracic Disease*, vol. 8, no. 1, 2016. [Online]. Available: http://jtd.amegroups.com/article/view/6353

[4] M. Brauer, G. Freedman, J. Frostad, A. van Donkelaar, R. V. Martin, F. Dentener *et al.*, "Ambient air pollution exposure estimation for the global burden of disease 2013," *Environmental Science & Technology*, vol. 50, no. 1, pp. 79–88, 2016, pMID: 26595236. [Online]. Available: https://doi.org/10.1021/acs.est.5b03709

[5] WHO, "Air pollution," 2018. [Online]. Available: http://www9.who.int/airpollution/ambient/health-impacts/en/

[6] I. C. Hanigan, R. A. Broome, M. Cope, J. S. Heyworth, J. R. Horsley, B. Jalaludin *et al.*, "The burden of mortality attributable to anthropogenic pm2.5 in australia, 2010-2016," *Environmental Epidemiology*, vol. 3, 2019.

[7] J. G. Su, G. S. Allen, P. J. Miller, and M. Brauer, "Spatial modeling of residential woodsmoke across a non-urban upstate new york region," *Air Quality, Atmosphere & Health*, vol. 6, pp. 85–94, 2011.

[8] A. M. Villalobos, F. Barraza, H. Jorquera, and J. J. Schauer, "Wood burning pollution in southern chile: Pm2.5 source apportionment using cmb and molecular markers," *Environmental Pollution*, vol. 225, pp. 514 – 523, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0269749116324666

[9] M. E. Quinteros, S. Lu, C. Blazquez, J. P. Cárdenas-R, X. Ossa, J.-M. Delgado-Saborit *et al.*, "Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in temuco, chile," *Atmospheric Environment*, vol. 200, pp. 40 – 49, 2019.

[10] M. del Medio Ambiente, "Decreto nº12 que establece norma primaria de calidad ambiental para material particulado fina respirable," 2011. [Online]. Available: https://www.leychile.cl/Navegar?idNorma=1025202

[11] P. Ruiz-Rudolph, F. Rubilar, M. Quinteros, K. Cayupi, S. Lu, R. Jimenez *et al.*, "Spatial distribution of particulate matter in winter nights in temuco, chile: a study of residential wood-burning impacts using mobile sampling," 08 2018.

[12] L. A. Díaz-Robles, J. S. Fu, A. Vergara-Fernández, P. Etcharren, L. N. Schiappacasse, G. D. Reed *et al.*, "Health risks caused by short term exposure to ultrafine particles generated by residential wood combustion: A case study of temuco, chile," *Environment International*, vol. 66, pp. 174–181, 2014.

[13] C.-X. Zhao, Y.-Q. Wang, Y.-J. Wang, H.-L. Zhang, and B.-Q. Zhao, "Temporal and spatial distribution of pm2.5 and pm10 pollution status and the correlation of particulate matters and meteorological factors during winter and spring in beijing," *Huan jing ke xue= Huanjing kexue*, vol. 35, no. 2, p. 418—427, February 2014.

[14] J. Lin, A. Zhang, W. Chen, and M. Lin, "Estimates of daily pm2.5 exposure in beijing using spatio-temporal kriging model," *Sustainability*, vol. 10, no. 8, 2018. [Online]. Available: https://www.mdpi.com/2071-1050/10/8/2772

[15] C. A. Garcia, P.-S. Yap, H.-Y. Park, and B. L. Weller, "Association of long-term pm2.5 exposure with mortality using different air pollution exposure models: impacts in rural and urban california," *International Journal of Environmental Health Research*, vol. 26, no. 2, pp. 145–157, 2016, pMID: 26184093. [Online]. Available: https://doi.org/10.1080/09603123.2015.1061113

[16] M. Eeftens, R. Beelen, K. de Hoogh, T. Bellander, G. Cesaroni, M. Cirach *et al.*, "Development of land use regression models for pm2.5, pm2.5 absorbance, pm10 and pmcoarse in 20 european study areas; results of the escape project," *Environmental Science & Technology*, vol. 46, no. 20, pp. 11 195–11 205, 2012, pMID: 22963366. [Online]. Available: https://doi.org/10.1021/es301948k

[17] L. Huang, C. Zhang, and J. Bi, "Development of land use regression models for pm2.5, so2, no2 and o3 in nanjing, china," *Environmental Research*, vol. 158, pp. 542 – 552, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0013935117312884

[18] M. Quinteros, C. Blazquez, F. Rosas, J. Cardenas, X. O. aand J. Delgado, R. Harrison *et al.*, "Development of land-use regression models for particulate matter due to residential wood burning in temuco, chile," *Environmental Epidemiology*, vol. 3, pp. 320–321, 2019.

[19] H.-C. Chuang, R.-H. Shie, C.-P. Chio, T.-H. Yuan, J.-H. Lee, and C.-C. Chan, "Cluster analysis of fine particulate matter (pm2.5) emissions and its bioreactivity in the vicinity of a petrochemical complex," *Environmental Pollution*, vol. 236, pp. 591 – 597, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0269749117343646

[20] S. Mahajan, H.-M. Liu, T.-C. Tsai, and L.-J. Chen, "Improving the accuracy and efficiency of pm2.5 forecast service using cluster-based hybrid neural network model," *IEEE Access*, vol. 6, pp. 19 193–19 204, 2018.

[21] Z. Chen, D. Chen, X. Xie, J. Cai, Y. Zhuang, N. Cheng *et al.*, "Spatial self-aggregation effects and national division of city-level pm2.5 concentrations in china based on spatio-temporal clustering," *Journal of Cleaner Production*, vol. 207, pp. 875 – 881, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959652618330956

[22] A. Peeters, M. Zude, J. Käthner, M. Ünlü, R. Kanber, A. Hetzroni *et al.*, "Getis-ord's hot- and cold-spot statistics as a basis for multivariate spatial clustering of orchard tree data," *Comput. Electron. Agric.*, vol. 111, no. C, p. 140–150, Feb. 2015. [Online]. Available: https://doi.org/10.1016/j.compag.2014.12.011

[23] L. Wang, H. Fan, and T. Gong, "The sales behavior analysis and precise marketing recommendations of fmcg retails based on geography methods," 2017. [Online]. Available: https://doi.org/10.20944/preprints201711.0115.v1

[24] L. Scrucca, "Clustering multivariate spatial data based on local measures of spatial autocorrelation," Université di Perugia, Dipartimento Economia, Finanza e Statistica, Quaderni del Dipartimento di Economia, Finanza e Statistica, Tech. Rep., 01 2005.

[25] N. Ohana-Levi, I. Bahat, A. Peeters, A. Shtein, Y. Netzer, Y. Cohen *et al.*, "A weighted multivariate spatial clustering model to determine irrigation management zones," *Computers and Electronics in Agriculture*, vol. 162, pp. 719 – 731, 2019.

[26] S. Y. J. Praetyo, K. D. Hartomo, B. H. Simanjuntak, and D. W. Candra, "Mitigation & identification for local aridity, based of vegetation indices combined with spatial statistics & clustering k means," *Journal of Physics: Conference Series*, vol. 1235, pp. 12–28, jun 2019.

[27] X. Wu, Y. Ding, S. Zhou, and Y. Tan, "Temporal characteristic and source analysis of pm2.5 in the most polluted city agglomeration of china," *Atmospheric Pollution Research*, vol. 9, no. 6, pp. 1221 – 1230, 2018.

[28] I. . Instituto Nacional de Estadística, "Censo 2017." [Online]. Available: https://www.censo2017.cl/

[29] B. Biblioteca del Congreso Nacional de Chile, "Reporte estadísticos comunales, 2017," 2017. [Online]. Available: https://reportescomunales.bcn.cl/2017/index.php/Temuco

[30] M. Hatzopoulou, M. F. Valois, I. Levy, C. Mihele, G. Lu, S. Bagg *et al.*, "Robustness of land-use regression models developed from mobile air pollutant measurements," *Environmental Science & Technology*, vol. 51, no. 7, pp. 3938–3947, 2017, pMID: 28241115. [Online]. Available: https://doi.org/10.1021/acs.est.7b00366

[31] J. V. den Bossche, J. Peters, J. Verwaeren, D. Botteldooren, J. Theunis, and B. D. Baets, "Mobile monitoring for mapping spatial variation in urban air quality: Development and validation of a methodology based on an extensive dataset," *Atmospheric Environment*, vol. 105, pp. 148 – 161, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1352231015000254

[32] W.-F. Ye, Z.-Y. Ma, and X.-Z. Ha, "Spatial-temporal patterns of pm2.5 concentrations for 338 chinese cities," *Science of The Total Environment*, vol. 631-632, pp. 524 – 533, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0048969718308118

[33] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[34] G. D. Silcox, K. E. Kelly, E. T. Crosman, C. D. Whiteman, and B. L. Allen, "Wintertime pm2.5 concentrations during persistent, multi-day cold-air pools in a mountain valley," *Atmospheric Environment*, vol. 46, pp. 17–24, 2012.

[35] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243 – 256, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S003132031200338X