Machine Learning Algorithms for Statistical Spatial Analysis of Waste Collection in an Italian Region

Gabriella Schoier(✉) and Giuseppe Borruso

DEAMS – Department of Economic, Business, Mathematic and Statistical Sciences "Bruno de Finetti", University of Trieste, Valerio, 4/1 – 34127, Trieste, Italy
{gabriella.schoier,giuseppe.borruso}@deams.units.it

Abstract. The aim of this paper is to consider machine learning algorithms to analyze the problem of the waste collection of an Italian region. This paper aims to thoroughly analyze data relating to separate waste collection in the Veneto region (Italy) at municipality level for the year 2023. Different variables have been considered.

After solving the problem of cleaning data a descriptive analysis has been done both with a graphical analysis using QGIS software and a statistical analysis using SAS software.

At this point some different machine learning methods has been applied using SAS in particular we used a principal component analysis to reduce the complexity of the phenomenon and then a non hierarchical cluster analysis to use a descriptive method.

The result indicates a constant improvement in the differentiated waste collection in the Veneto region.

Keywords: Machine Learning · Spatial Data Mining · Multivariate Methods · Waste Collection

1 Introduction

In the last decades one of the main problem has been how to extract information from data in our case we consider spatial data [10][1].

Big data analytics could provide opportunities to develop new knowledge to reshape our understanding of different fields and to support decision-making [7].

A huge amount of data i.e. big data have been collected by private and public organizations (see eg. [11]).

Moreover AI is becoming a part of our life and is conditioning our job, private life and social relationships and nowadays is strictly in conjunction with decisions in almost all our life [11].

[1] We are grateful to doctor Jacopo Vergine for the elaborations and the comments.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16761, pp. 557–567, 2027.
https://doi.org/10.1007/978-3-032-30527-5_36

This analysis is connected with the problem of sustainability at urban level that is a crucial economical social and political objective. To analyze the notion of sustainable urban life and to choose appropriate indicators it is necessary to choose appropriate indicators enabling its measurement, this is what the region Veneto try to do [5, 8].

The aim of this paper is to try to explain the development of the waste collection in an Italian region i.e. Veneto region which is in the North-East of Italy, its most known city is Venezia.

Based on the most recent data available, Veneto remains one of Italy's leading waste management regions, with high rates of separate waste collection and a strong focus on the circular economy [1, 4, 5].

The data are collected in the year 2023 by Arpa Veneto (see [1]) and regards different type of waste.

After a cleaning of the data a spatial descriptive statistical analysis as been performed to individuate the main characteristics of the waste different collection.

At this point we consider the possibility of obtain a piece of information by reducing the number of variables using principal component analysis (PCA) (see e.g. [2, 3]).

The results are encouraging as it is possible to give a label to the new variables.

On the base of these new variables a non hierarchical cluster analysis has been performed. The results can be useful to improve different policies.

1.1 The Data and the Variables

The dataset regards the waste collection in the Veneto region in the year 2023.

It contains detailed data relating to the production of municipal waste considering different aspects of waste collection. This dataset contains geographical, graphic and demographic information such as the name of the basin, the name of the municipality and the population, furthermore, for each municipality there are the different types of waste in kilograms. From these data, it is possible to obtain a comparative measure of the waste produced per person in each municipality, allowing for the analysis of waste management efficiency and differences between different locations.

The variables present in the dataset are:

- ISTAT: ISTAT numerical code that uniquely identifies the municipality;
- Territorial basin of the region Veneto for instance Padova Sud, Padova Centro etc.
- Municipality: name of the municipality to which the data refers;
- Population (Popolazione): number of inhabitants residing in the municipality FORSU (organic fraction of urban solid waste) collected
- Green (Verde): kg per capita of plant waste (pruning, grass, mowing) collected
- Glass (Vetro): kg per capita of glass collected in differentiated mode
- Paper and Cardboard: (Carta e Cartone) kg per capita of paper collected
- Plastic (Plastica): kg per capita of plastic collected
- Metal (Metallo): kg per capita of metal waste collected
- Wood (Legno): kg per capita of wooden material collected
- Multimaterial (Multimateriale): kg per capita of wooden material collected

- WEEE (Waste from Electrical and Electronic Equipment): kg per capita of waste from electrical and electronic equipment collected
- Textiles (Tessili) kg per capita collected
- Other Recoverable (Altro Recuperabile) kg per capita collected
- Special wastes (Rifiuti Particolari) kg per capita collected
- Bulky wastes (Ingombrante) kg per capita collected
- Sweeping (Spazzamento) kg per capita collected
- Residue (Residuo) kg per capita collected
- RU (Urban Waste) (Rifiuto Totale): kg per capita of municipal waste generated which includes both the differentiated fraction (RD) and the non-differentiated fraction;
- RD (Differentiated Collection) (Rifiuti Differenziati): kg per capita of waste collected differently, calculated as the sum of the individual fractions, in the municipality
- RUND (Non-Differentiated Urban Waste) (Rifiuti non Differenziati): quantity of indierenziated waste (dry residue) collected in the municipality, i.e. the part of RU excluding the RD.
- Percentage of separate waste collection (% Raccolta Differenziata) kg per capita collected
- Domestic users (Utenze Domestiche) represents the housing units registered in the waste collection service (TARI) as a residential use.

After importing the Excel file into the QGIS software, the data set variables were merged into the spatial layer using the Spatial Joins function.

These modifications allowed us to obtain a dataset ready for the subsequent analysis phases and a shapefile suitable for an exploratory spatial analysis. From this point on, it is also possible to generate geographic visualizations of the distributions of the common seventeen variables using the QGIS Symbology function. The following graphs collected in Fig. 1 illustrate the positive trend in the percentage of separate waste collection in recent years, obtained by downloading and displaying shapefiles for previous years.

The trend reflects effective environmental policies, integrated waste management and a growing awareness of the population, which allowed the Veneto Region, with a total percentage of 77.6%, to exceed both the national average (65.2%) and that of Northern Italy (71.2%) in 2023 [1].

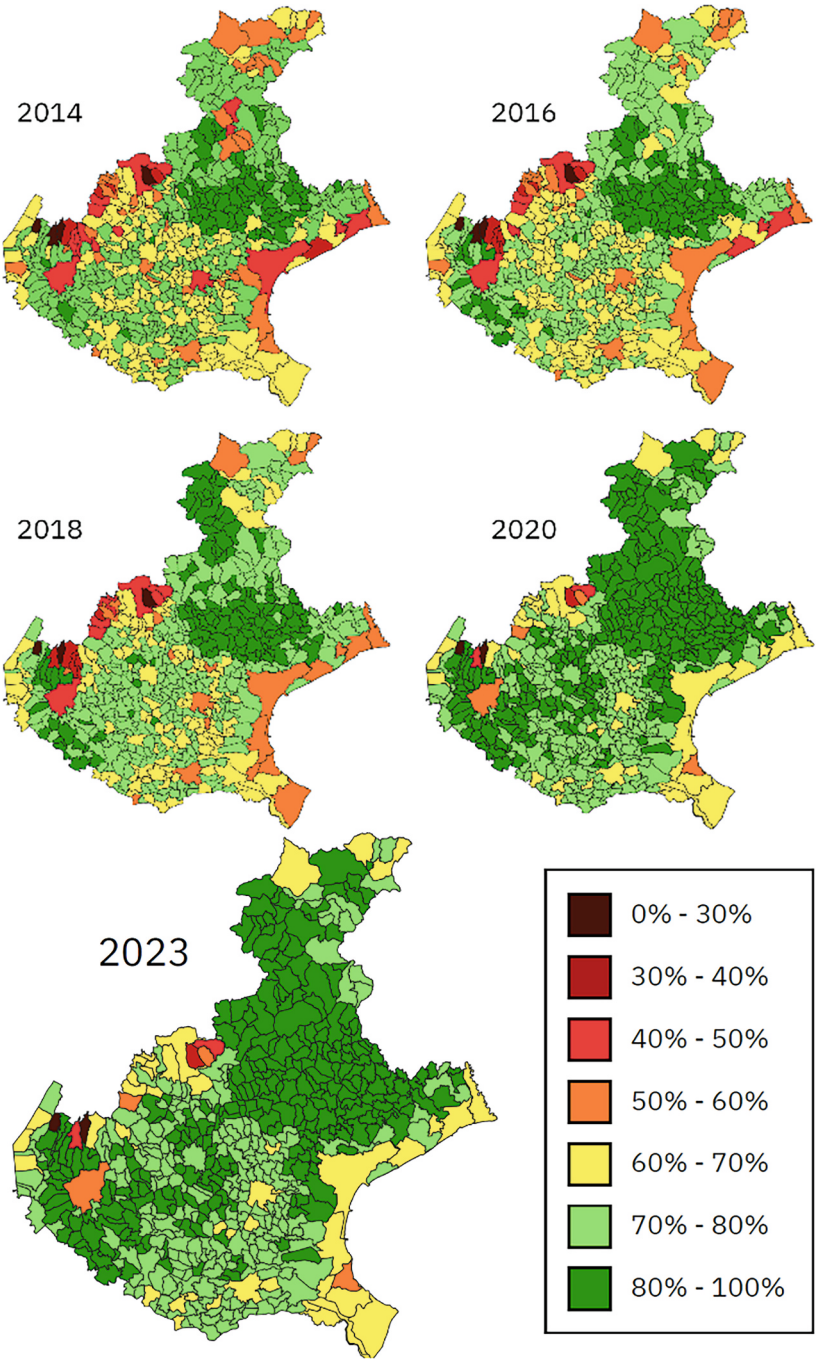

Fig. 1. Maps representation in different years of percentage of differentiated waste

1.2 A Descriptive Spatial Statistical Analysis

In the following we present some descriptive statistical analysis regarding the collected variables (Table 1).

Table 1. Descriptive Statistics

Variable	N	Mean	Median	Std Dev	Minimum	Maximum	1st Pctl	99th Pctl	IQR	Skewness	Kurtosis
Popolazione	563	8647.57	4852	19270.37	126	257213	320	49985	7010	9.95	116.2
FORSU	559	80.16	76.72	25.91	25.52	236.71	32.38	184.76	23.11	1.84	6.34
Verde	552	62.72	55.36	47.18	0.33	514.38	3.63	226.01	50	3.28	22.36
Vetro	531	35.64	36.82	23.88	0.02	207.06	0.27	129.45	11.77	2.33	13.85
Carta e Cartone	562	54.27	50.61	16.77	18.36	160.87	32.39	138.89	14.41	2.79	12.41
Plastica	344	8.63	4.45	12.52	0.01	66.23	0.1	60.09	5.98	2.59	6.28
Metalli	504	6.8	5.22	5.42	0	28.58	0.27	22.78	5.91	1.49	1.75
Multimateriale	553	51.01	40.02	29.24	0.04	205.4	0.36	148.66	26.15	1.71	3.93
RAEE	555	5.88	5.48	2.46	0.02	18.76	1.72	12.17	2.62	1.4	3.73
Tessili	510	3.51	3.25	2.77	0	41	0.12	10.35	2.83	5.42	65.9
Altro Recuperabile	506	21.93	20.39	11.14	0.04	101.84	0.43	52.29	10.65	1.52	6.06
Rifiuti Particolari	559	1.74	1.69	0.83	0.08	5.01	0.31	4.37	0.97	0.97	1.71
Ingombrante	562	18.85	16.92	9.68	1.23	95.9	4.84	54.53	9.78	2.59	12.81
Spazzamento	479	11.47	7.74	12.7	0.28	144.84	0.61	66.5	6.91	4.39	30.87
Residuo	563	89.55	76.87	68.65	23.47	636.95	26.37	434.28	45.39	3.9	20.55
Rifiuto Totale	563	438.89	411.42	155.02	227.75	1685.4	276.42	1305.66	109.22	4.34	25.73
Rifiuti Differenziati	563	353.58	335.38	104.9	161.67	1181.46	188.32	887.85	77.52	3.76	21.91
Rifiuti Indifferenziati	563	85.3	71.57	68.09	19.56	626.74	23.54	417.87	44.81	3.89	20.61
% Raccolta Differenziata	563	0.82	0.82	0.08	0.28	0.95	0.56	0.93	0.1	-1.75	7.73
Utenze Domestiche	522	0.11	0.08	0.12	0	2.1	0	0.44	0.1	8.47	125.72

Although some variations between municipalities fall within the physiological variability expected for this type of indicators, the analysis found marked variability in the distributions of observed values, with standard deviations suggesting, in most cases, strong territorial heterogeneity. This phenomenon is also reflected in the production of total waste, which is characterized by great variability, extreme values that are very far from each other and high values of asymmetry and kurtosis. These elements suggest a notable diversification in local waste gestion behaviors and strategies, with some municipalities deviating significantly from the average. For many variables, moreover, interquartile range (IQR) is high, indicating that within the central 50% of the data there is considerable variability.

The skewness of the distribution is positive for almost all variables, suggesting longer tails to the right due to the presence of a few municipalities with values much higher than the average. The only exception is the percentage of separate waste collection, with negative asymmetry, which indicates that most municipalities are concentrated towards high values. Variables such as Textiles, Sweeping, and Total Waste also exhibit very high kurtosis, suggesting heavy tails.

Other variables, such as WEEE, Metals and Particular Waste present a more "moderate" kurtosis, indicating a distribution closer to a normal one.

We proceeded with a univariate analysis of the distribution of all numeric variables, using the univariate procedure of SAS and the Symbology function of QGIS. Below are the most relevant visualizations and only a few particularly significant variables are analyzed (Fig. 2).

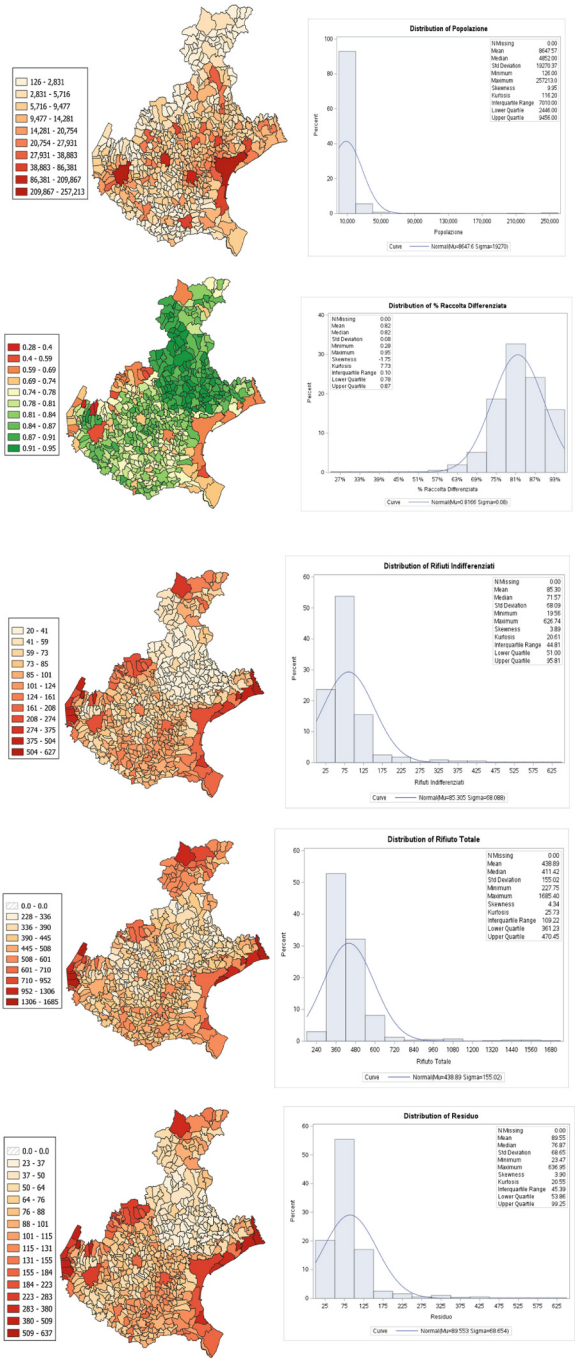

Fig. 2. Maps representation of different types of waste

As already highlighted, the dataset shows considerable heterogeneity between municipalities.

The main variables show highly skewed distributions, with mean values often unrepresentative due to the presence of outliers. Although these are relationships that should be formally analyzed, it is immediately evident, from the visualization of geographical distributions, that the variability in values is often explained by the presence of tourists, morphological factors of the territory, population density and the position of industrial zones.

The distribution of Total Waste Per Capita, for example, reflects population density and tourism. The strong rightward skewness and the heavy tails of the distribution indicate the presence of municipalities with a much higher per capita waste production than the regional average i.e. large cities and tourist resort in which the presence of hotel, restaurants and commercial activities produces waste flows that cannot be assigned to the resident population. For instance the municipalities around Lake Garda, Venice and the coastal areas show very high values of waste.

The percentage of Differentiated Collection has a negative skewness, with most values concentrated at the top of the distribution. The few negative exceptions are related to large cities, which probably present critical issues related to population density, or organization and tourist resorts (in which external influx generates large quantities of waste. Finally, lagoon areas and small mountain and isolated municipalities show difficulties related to logistics which complicates waste management.

The areas with the highest values of residual Waste and undifferentiated waste per capita correspond to tourist cities and, to a lesser extent, to industrial areas and areas where collection services are probably less efficient. Low values, however, correspond to the central Veneto region and part of the Belluno area, where municipalities show good performance thanks to efficient door-to-door collection systems and extensive service coverage.

2 Principal Component Analysis

2.1 Some Results

Principal component analysis (PCA) is a widely used unsupervised dimensionality reduction technique that simplifies complex, high-dimensional datasets into a smaller set of uncorrelated variables the so called principal components trying to retain the maximum variance explained by the original variables.

This methods allows to reduce the dimensionality of the dataset and identify underlying patterns or structures in order to understand which variables explain most of the variability.

In our analysis, all variables relating to the product categories of differentiated and undifferentiated waste were selected. It was decided to exclude the overall totals, since, being sums of the components already present, they are linearly dependent or highly colli-near with the variables used. The variable “Population” has been also excluded, as it is just comprehensive in per capita calculation of the indicators.

The output of the PCA includes several fundamental elements for interpretation. The covariance or correlation if we consider standardized variables matrix describe how the variables in the dataset depend on each other. This is essential in PCA, because

principal components are constructed as linear combinations of variables that explain the maximum possible variance in the dataset and are uncorrelated. In the matrix generated by SAS Studio, strong covariances are noted between some variables, such as Paper.

Conversely, variables such as Green and Textiles show very low or negative covariances, indicating different dynamics (Table 2).

Table 2. Eigenvalues

Eigenvalue	Eigenvalue value	Proportion	Cumulative
1	4.61	0.38	0.38
2	2.74	0.23	0.61
3	1.05	0.09	0.70
4	0.84	0.07	0.77
5	0.66	0.05	0.82

The table shows the eigenvalues of the covariance matrix, each representing the amount of variance explained by each principal component. A high eigenvalue indicates that the corresponding component can explain a significant share of the overall variability of the data set. The eigenvalues indicate that the first component (Prin1) alone explains 38.4% of the variance, the first two components together explain 61.3%, and the first three components reach 70.1%, thus reaching an already considerable threshold. The first five components finally explain l'82.6%, a high share that indicates a good explanation of the initial phenomenon.

Eigenvectors indicate the contribution of each original variable to each component. The values therefore show how the components are built as linear combinations of the original variables moreover these new variables are uncorrelated.

The higher the absolute value of the coefficient, the greater the weight that variable has on the component.

The important thing is if we can understand the meaning of the new variables otherwise these new variables can not be used in subsequent analysis.

In this case:

- Prin1 is characterized by high and positive coefficients for most of the variables, in particular Multimaterial (0.378), WEEE (0.370), Paper and Cardboard (0.354), Recoverable Other (0.349), Metals (0.330), Sweeping, FORSU and Particular Waste. This component effectively synthesizes the overall intensity of separate collection. It could be labelled as *Component of the Complex Differentiated Collection.*
- Prin2 shows the highest coefficients on Green (0.446) and Residue (0.456), while Particular Waste and WEEE show negative values. This dimension seems to reflect a dichotomy between organic or plant-based waste and more technological waste. It can be defined as *Environmental-Organic Component*.
- Prin3 is dominated by the Textiles variable (0.766), the only one to have a high loading, while all the others show values close to zero or negative. This component appears to identify a specific variability related to a rare and inhomogeneous fraction. It can be interpreted as a *Component of Textile Specificity.*

- Prin4 shows high values on Textiles (0.566) and Bulky (0.507), which can be repaired and could indicate occasional or non-systematic waste. It could be seen as a *Component of Special Waste.*
- Prin5 has the highest loads on Sweeping (0.459), Bulky (0.537), Residue (0.252), with opposite directions for other variables. It could represent an *Urban Structural Component*, linked to waste management in urban areas. These labels summarize the structural and territorial significance of the components, and could also be used in subsequent analyses (e.g. regression, clustering, or classification), as well as in spatial mapping for the identification of regional patterns.

The choice to keep up to 5 components is supported by the cumulative curve that consider the explained variance, but for a synthetic and visual analysis, the first 2 or 3 are the most interpretable (Figs. 3 and 4).

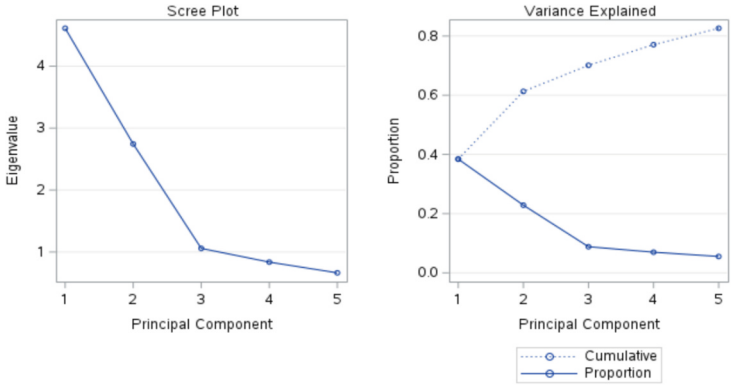

Fig. 3. Scree Plot and Variance Explained

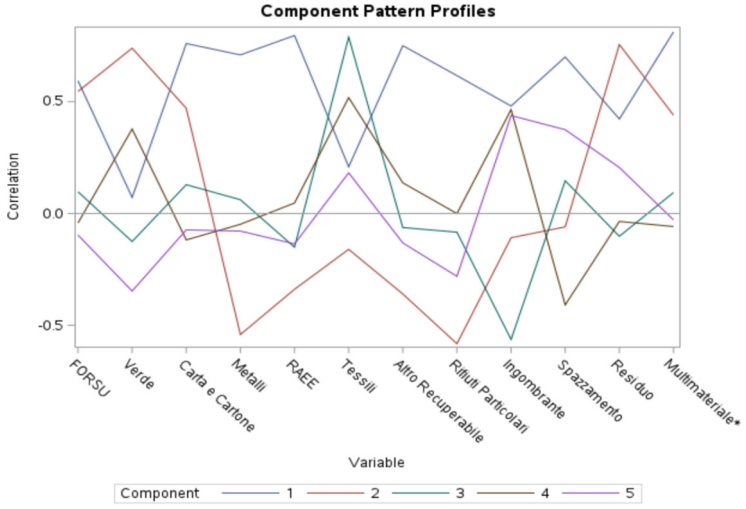

Fig. 4. Component Pattern Profiles

The Component Pattern Profiles graph shows the correlation between each original variable and the first five main components, supporting the interpretation of the new variables.

The colored lines represent the trend of the eigenvectors for each component, highlighting which wastes contribute most to each latent dimension.

The PCA results are overall good and indicate that the analysis is useful, especially considering the socio-ambiental context of the study. The analysis effectively reduced the complexity of the dataset by synthesizing twelve original variables into components components that explain over'82% of the variance.

In particular, the first two components explain over 61% of the total variance, the first three reach 70%, and the first five exceed'82%, allowing for a compact and faithful representation of the data.

The presence of a sharp "elbow" within the scree plot between the second and third components suggests that the first components capture the dominant information in the dataset. The eigenvectors of the principal components finally show a coherent structure: some variables contribute strongly to the first components, suggesting non-random structured patterns, which makes it possible to attribute meaning to the new components as seen.

3 Cluster Analysis

On the base of the results of the PCA a non-hierarchical cluster analysis has been performed [6]. In more details we used a k-means method implemented on SAS platform [9].

The first cluster contains 5 observations among which Cortina d'Ampezzo, its standard deviation is equal to 1,533 and is the highest, so it could be said that this cluster is the least compact. The maximum distance between a point and the center of the cluster is 3.65.

The second cluster contains 23 observations.

The third cluster contains 202 observations it contains 35.88%of the observations.

The fourth cluster contains more than half of the observations, has a frequency of 326 or a percentage of 57.90%. This cluster has a standard deviation less than 1 and is therefore more compact than the other clusters. There is a satisfactory separation since the sum of the r-squares is slightly high.

The fifth cluster contains 7 observations,

As there are a lot of municipalities we just have a look to the observations of the provinces of Belluno and Vicenza.

The province of Belluno has 61 municipalities in particular the municipality of "Cortina d'Ampezzo" is assigned to the first cluster and the remaining 60 municipalities are assigned to the third cluster, the municipality of "Auronzo di Cadore" is the furthest from the center of the cluster since it has the maximum distance between the centroids which is equal to 3.942, the municipality of "Vigo di Cadore" is the closest from the center of the cluster since it has the minimum distance between the centroids which is equal to 0.472, the municipality of "Colle Santa Lucia" is the municipality with the

highest metallurgical index, it is equal to 4.408, the municipality of "Borca di Cadore" is the municipality with the highest organic degradability index, it is equal to 1,962.

The province of Vicenza has 90 municipalities, 8 municipalities are assigned to the second cluster, 20 municipalities to the third cluster and 62 municipalities to the fourth cluster. The municipality of "Calvene" is the furthest from the center of the cluster since it has the maximum distance between the centroids which is equal to 2,733. The municipality of "Orgiano" is the closest from the center of the fourth cluster since it has the minimum distance between the centroids which is equal to 0.552.

4 Conclusions

In this paper we have analyzed data from waste collection in an Italian region i.e. Veneto region, we have also illustrated processes of extraction of information from this kind of data.

The analysis can be extended in various ways, such as applying clustering algorithms on PCA scores, or using the first components as input to predictive models, including classification decision trees. In our case after an application of a principal component analysis we have consider a non hierarchical cluster analysis.

The results show that in the year the citizens' awareness regarding separate collection has increased over time. This has led to an improvement in waste storage conditions.

Moreover the subsequent non hyerachical cluster analysis has show the differences in the municipalities. This is due to the diversity of the territory from coast to mountain and to the presence of art cities.

References

1. Arpa Veneto (2026). https://www.arpa.veneto.it/temi-ambientali/rifiuti/rifiuti-urbani/rifiuti-urbani-2023
2. Bolasco S.: Analisi Multidimensionale dei Dati. Metodi, strategie e criteri d'interpretazione, Carrocci editore, Roma (2005)
3. Gareth J., Witten D., Hastie T., Tibshirani R.: Introduzione all'apprendimento statistico. In: Salini, S., Gaito, S., Boracchi, P., Ambrogi, F., Manzi, G., Biganzoli, E. (eds.) Piccin Editore (2021)
4. ISTAT: Il Censimento permanente della popolazione in Veneto Anno 2023 (2025)
5. ISTAT: Il Benessere Equo e Sostenibile dei Territori: Veneto 2025 (2025). https://www.istat.it/wp-content/uploads/2025/12/BesT2025_Veneto.pdf
6. Jan, A.K.: Data Clustering. 50 years beyond K-means. Pattern Recogn. Lett. 651–666 (2010)
7. Mayer-Schonberger, V., Cukier, K.: Big Data: A Revolution That Will Transform How We Live, Work, and Think. Mariner Books, Boston (2013)
8. Monte, A., Schoier, G.: A Multivariate statistical analysis of equitable and sustainable well-being over time. Soc. Indicat. Res. **161**(2–3), 735–750 (2020)
9. SAS on Demand. https://welcome.oda.sas.com/. Accessed 30 Mar 2026
10. Schoier, G., Borruso, G.: A methodology for dealing with spatial big data. Int. J. Bus. Intell. Data Min. **12**(1), 1–13 (2017)
11. Software Testing Help: Data Mining Vs Machine Learning Vs Artificial Intelligence Vs Deep Learning. http://www.intelligenzaartificiale.it/data-mining/. Accessed 30 Mar 2026