Behind the Gates: Deciphering Milan's Area C Traffic DNA

Ca' Foscari University of Venice CM90
Computer Science and Information Technology
CM0471 - Statistical Inference Learning
Author: Gabriele Pilotto – 902388

Introduction

In 2012, Comune di Milano introduced **Area C** which is a combined congestion charge and Low Emission Zone (ZTL) covering the area called "*Cerchia dei Bastioni*". Its aim is to reduce traffic, promote public transport and lower pollution levels by limiting access to Milan's city center. The Area C is delineated by **ANPR (Automatic Number Plate Recognition) "gates"** monitoring the vehicles entrance 24/7 and logging aggregated information about them.

The **goal of this project** is to **understand more about the Area C users** using prediction and inference techniques.

1.1. Area C dataset

The data used for this project are publicly available at https://dati.comune.milano.it/dataset. The dataset is provided by Comune di Milano while its maintenance and review is delegated to AMAT. For this work, the latest available data are utilized: from January 2024 to November 2024.

The columns representing the dataset are listed and commented on in the `01-data_preparation` document. Overall, the dataset contains13 columns and about 12M rows, each representing a 30-minutes aggregation window.

2. Data preparation

Before starting the analysis, a data review was necessary, including removing and cleaning variables, aggregating information, and performing feature engineering for subsequent modules. Following a preliminary database analysis, the modifications made are:

- Removal of the fap and moto variables, as they were not complete or useful for searching
- Remapping `tipologia_alimentazione` to remove unused options and merge similar fuel types.
- Removal of duplicates generated by the two changes.
- Type conversion and labeling to improve readability and correct information management, especially during classification operations.
- Breakdown of the `dataora` variable into subcomponents and subsequent removal of duplicates.
- Feature engineering for the following variables:
    - `holidays_2024` containing the official holidays in Italy.

    - `is_weekend`, `is_holiday`, `is_rush_hour`, and `is_post_closure` were created starting from the timestamp breakdown. Some of these variables will not be used extensively but were created as a basis for possible future work.
    - To create the `is_pollutant` variable, information about the policies adopted, the date, and the fuel type were combined. The type of check to determine whether a vehicle is polluting or not considers the October 2024 regulatory update.
- Creation of a new dataset for hourly traffic is useful for quantitative analysis.
- Creation of new simple features containing the GPS coordinates of the gates.

3. Analysis

The conducted analyses are both qualitative and quantitative. The goal of deciphering the Area C traffic DNA was achieved combining multiple research questions and approaches. For under performative models, alternative methods were tried until a proper result was reached. However, some questions could not be resolved with sufficient accuracy; the difficulties encountered are documented in this report and in the RMarkdown files where a more detailed explanation is offered.

3.1. Milan's habits

The first research focus is based on understanding the habits of Milan's citizens. This includes identifying the traffic patterns, the influence of office hours, traffic-increasing context and geographic traffic distribution.

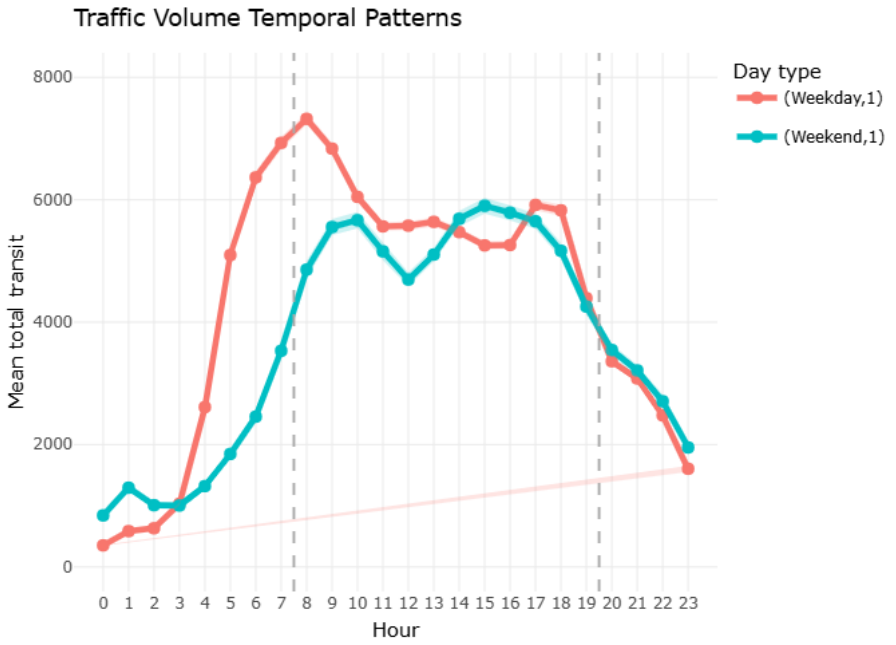
3.1.1. Traffic during Rush Hours

Research question	Does the rush hour correspond to office hours? Are there some anomalous peaks?
Dataset	Full dataset `df_features.rds`
Model	Exploratory Data Analisys
File	`02-exploratory_data_analysis.rmd`

To understand if the traffic pattern is influenced by the offices' working hours, a division is made between weekdays and weekend (since on Saturday and Sunday offices are mostly closed thus the effect is reduced or absent). If the traffic is influenced by the offices, a "M" shape is expected with peaks at 08:00 and 17:00.

The traffic dynamics of Area C reveal a two-headed rush hour that bifurcates weekdays morning congestion into two distinct phases:

- **Service Rush Hour** (04:00–07:30), consists of logistics vehicles entering early to avoid toll
- **Office Rush Hour** (08:00–09:30), the originally expected rush hour peak



The expected "M" shape turned out to be asymmetric: the afternoon peak is less pronounced. This happens because the exit flows are untolled and temporally dispersed, unlike the cost-sensitive morning entry where some vehicles access the gates before 07:30.

The graph draws also the weekend "lazy Milan" characterized by a late awakening where the traffic starts to intensify after 10:00. Intense traffic volume in the afternoon (shopping) and the nightlife keeping high traffic density until 02:00, confirming the habit shift.

3.1.2. Traffic-increasing Context

Research question	What leads to more traffic?
Dataset	80/20 split dataset `df_hourly.rds`
Model	Multiple Linear Regression, Polynomial Regression, Iterative Polynomial Regression, GAM
File	`03-quantitative_data_analysis.rmd`

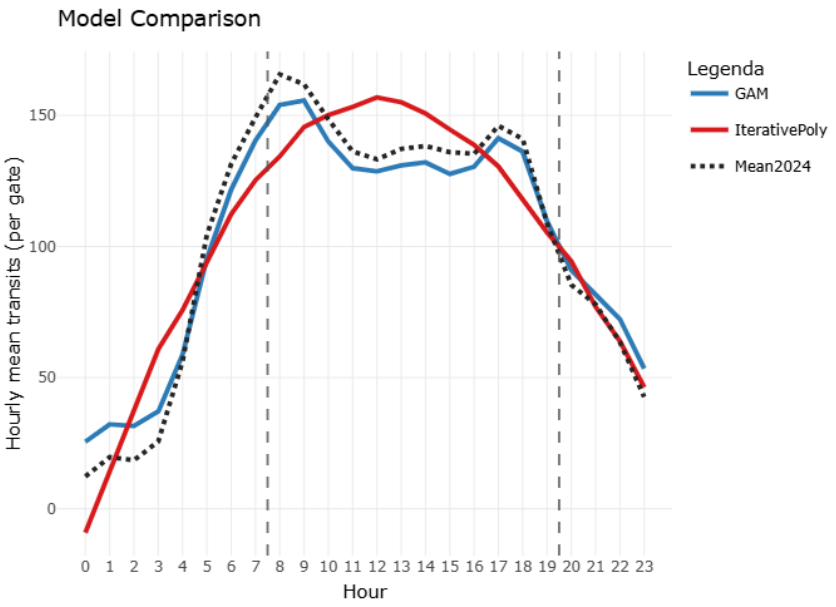
Answering this question requires multiple statistical approaches, from a simple linear one to the GAM spline to better capture the cyclic nature of urban mobility.

Method 1: Initial attempts using *Multiple Linear Regression* proved inadequate, explaining only 4% of traffic variability. The failure stems from the model's inability to account for the "M" shaped traffic pattern seen in the EDA.

Method 2: To address this non-linearity, a *Polynomial Regression* was implemented. By introducing quadratic and cubic terms for the hour variable, the model can now represent morning and afternoon peaks, improving the adjusted R^2 to 18% and high residual standard error.

Method 3: This led to the *Polynomial Interaction Model*, allowing the hour predictor to interact with `day_of_week` and gate `label`. This synergy significantly boosted the model's explanatory power to 72%, confirming that traffic shapes fluctuate based on both time and location. However, linear models struggle with extreme outliers, and the variance was increasing alongside with traffic volume.

Method 4: The final *Generalized Additive Model (GAM)* was adopted to provide maximum flexibility. By utilizing Negative Binomial distribution, the GAM effectively handled the spikes and high variance that simpler models ignore.



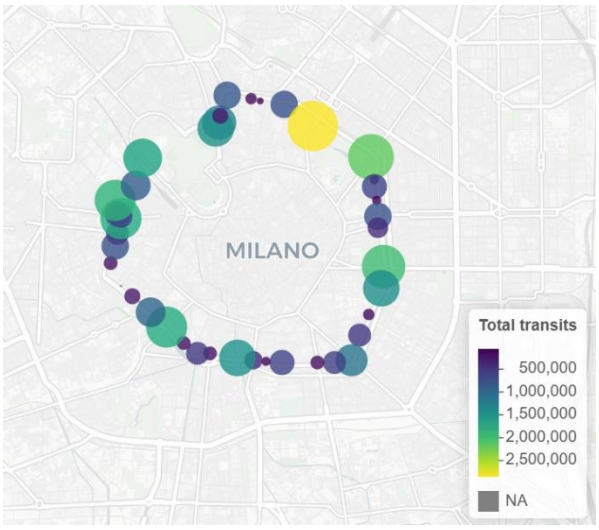
Using a spline with 20 degrees of freedom, the GAM achieved an R^2_{adj} of 84% and explained 85.2% of the deviance. Good results may be achieved also with 18 grades of freedom, but the reduction may penalize the peaks identification.

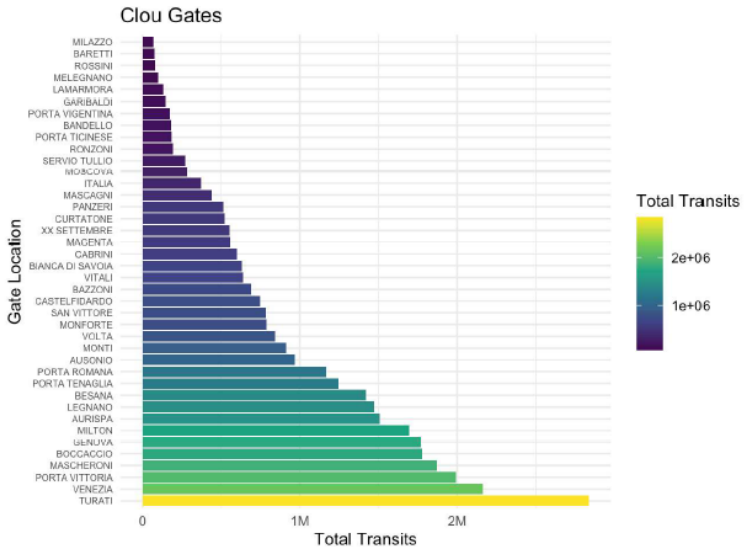
Based on those models it's safe to assume that the causes of more traffic are mostly social/working habits and environmental policies. However, the high residuals suggest that more predictors (not included in the dataset), such as weather, important events date and construction sites, should be included to increase the model accuracy and reduce errors.

3.1.3. Clou Gates

Research question	Which are the clou-gates?
Dataset	Full and balanced 80/20 split dataset `df_features.rds`
Model	Exploratory Data Analysis, Naive Bayes
File	`04-qualitative_data_analysis.rmd` `02-exploratory_data_analysis.rmd`

The analysis identifies the primary access gates to Area C by combining a geo-spatial volume analysis and a probabilistic behavioral model.





Method 1: Ranking gates by total volume revealed a high concentration of entries along the northern and eastern axes. *Turati* emerged as the primary corporate gate, tunneling traffic from Central Station and *Porta Nuova* toward the Fashion District. *Venezia* followed as the natural terminus for Corso

Buenos Aires, while *Porta Vittoria* (East) and *Boccaccio/Mascheroni* (West) handle high residential and airport-bound volumes.

Method 2: To move beyond raw numbers, a *Naive Bayes* model was utilized to determine the probability of gate selection based on resident status, vehicle type, and hour. However, the Naive Bayes approach yielded a 22% accuracy, largely due to high attraction of the Turati gate where the model over-assigns transits to the highest-volume gate. Due to low accuracy, we can't consider the model outcome as valid.

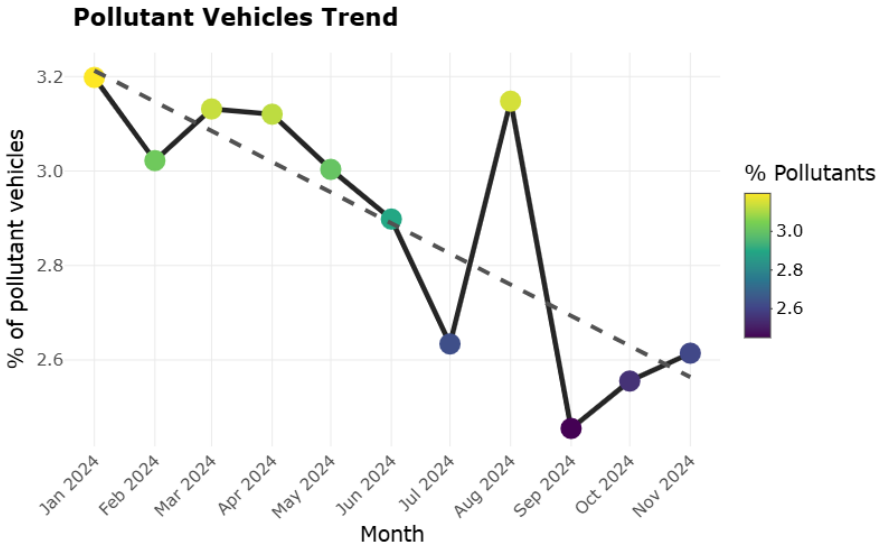
3.2. Environmental Aspects

Area C project was born to reduce smog and pollutant concentrations over the city center. This section's objective is to understand if (over the 11-month considered) the pollutant vehicles access is decreasing, especially considering new rules and restriction imposed after October 2024. To understand where Area C can improve, a context analysis is provided to understand where the most pollutant vehicles access the city.

3.2.1. Euro-rule Effectiveness

Research question	Is the "Euro-rule" working? Is the most-pollutant-vehicles traffic decreasing?
Dataset	Full dataset `df_features.rds`
Model	Exploratory Data Analysis, Linear Regression
File	`02-exploratory_data_analysis.rmd`

This analysis examines the downward trajectory of polluting vehicles within Area C, evaluating the statistical significance of legislative shifts and seasonal behavior.



Method 1: Exploratory Data Analysis reveals a consistent downward trend in the percentage of polluting vehicles over the 11-month study period. An exception must be made for august where holiday traffic or renovation works may have significantly improved the entrance.

Method 2: A linear regression model confirms this trajectory with a negative slope of -0.0021278, indicating a statistically significant, albeit gradual, reduction. The linear model effectively captures the direction of the trend, but its R^2 of 57% indicates that over 43% of the variability is driven by external factors.

A critical juncture occurred in October 2024, when restrictions were extended to include Euro 3 petrol vehicles. While this expanded definition of "*pollutant*" was expected to cause a sharp data spike, the results may suggest a seasonal effect rather than a penalty for older petrol cars.

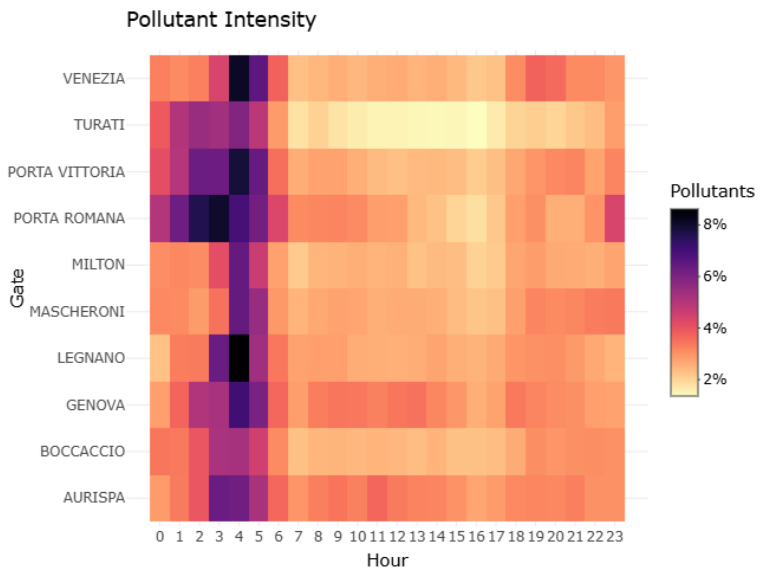
3.2.2. Pollutant Context

Research question	In what context do the more-pollutant-vehicles access the most?
Dataset	Balanced 80/20 split and full dataset `df_features.rds`
Model	Exploratory Data Analysis, Linear Discriminant Analysis
File	`02-exploratory_data_analysis.rmd` `04-qualitative_data_analysis.rmd`

Combining *Linear Discriminant Analysis* with the *Exploratory Data Analysis* provides a comprehensive view of the pollutant identity in Milan's city center, merging statistical drivers with spatial and temporal behaviors.

Model 1: LDA with accuracy, sensitivity, and specificity, all reaching approximately 80%**,** helps to identify *who* is polluting:

- Diesel**:** most pollutant vehicles are diesel-powered. Specifically, diesel vehicles with a classification below Euro 5 constitute 88.2% of all diesel transits
- Logistic**:** vehicles used for goods transport are 29.3% more likely to be classified as pollutant
- Time: pollutant vehicles enter Area C earlier linking pollution to early-morning logistics



Explorative Data Analysis: On the other hand, a heatmap confirms the supposition that may be made from the LDA results: logistics hubs (*Venezia, Porta Romana, Porta Vittoria*) exhibit the highest pollutant intensity. Their location connects the city center to eastern and southern quadrants rich in logistics warehouses, aligning with the LDA's finding that "Goods" classification is a major pollutant predictor.

3.3. User profiling

Last section is focused on creating an identikit of the Area C user. The analysis relies only on classification models since the goal is to identify a resident among all the transiting users.

3.3.1. Resident Discriminators

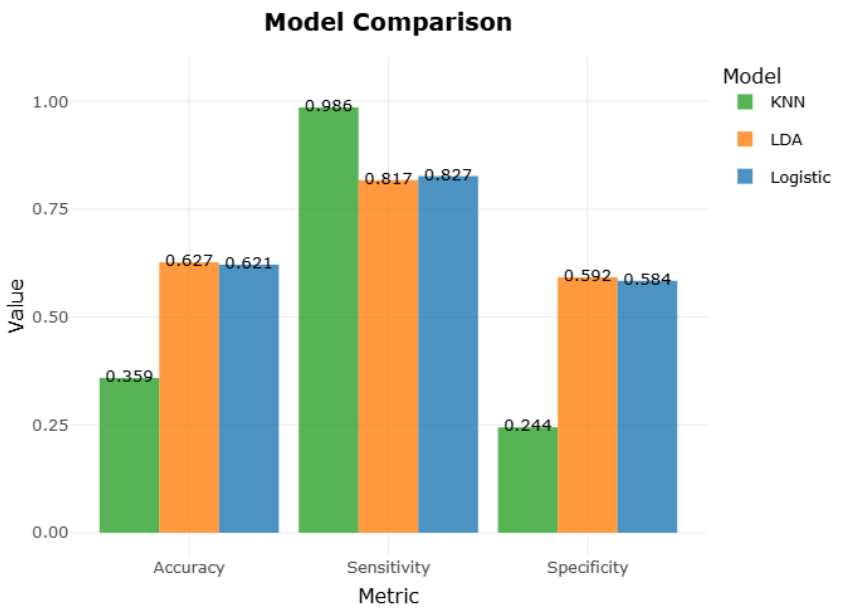
Research question	What does distinguish a resident? Is it the timestamp? The gate? The vehicle type?
Dataset	Balanced 80/20 split dataset `df_features.rds`
Model	Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbor, EDA
File	`02-exploratory_data_analysis.rmd` `04-qualitative_data_analysis.rmd`

Identify a "resident identikit" wasn't an easy task. Finding characteristics that discriminate residents from non-residents accessing Area C, due to the heterogeneity of urban mobility patterns in Milan, has requires multiple classification approaches.

Model 1: *Logistic Regression* was used to determine the odds of a user being a resident based on specific predictors. This model serves as a strong baseline, achieving an accuracy of 62.1% and a high sensitivity of 82.7% in identifying actual residents. Being a resident is statistically less probable by default. However, the probability increases significantly if the user drives a petrol-powered private car. Residents show a strong preference for petrol engines (Odds Ratio: 2.81), while electric vehicles are surprisingly rare among this group. This suggests that residents likely own traditional cars for short trips or high-performance vehicles, which are prevalent in Milan's exclusive city center.

Model 2: The *LDA* model projects data to maximize the separation between residents and non-residents, achieving the highest overall accuracy at 62.7%. While confirming the results achieved with logistic regression, LDA shows that, despite the lack of electric vehicle adoption, residents are statistically more eco-friendly than visitors, with a significantly lower probability of driving the most polluting vehicle categories (0.02 vs 0.09).

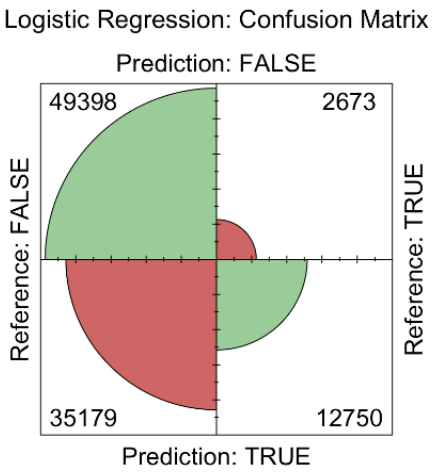
Model 3: Trying to improve the accuracy, The *KNN* approach (specifically *KKNN* to handle data ties) checks the proximity of transits without making linear assumptions. However, this model proved to be ineffective for general classification in this context. KNN achieved near-perfect sensitivity (98.6%), its specificity was critically low (24.4%). Due to the high density of transits during peak hours, the model



drastically overfit, categorizing almost every user as a resident because they are spatially and temporally surrounded by them.

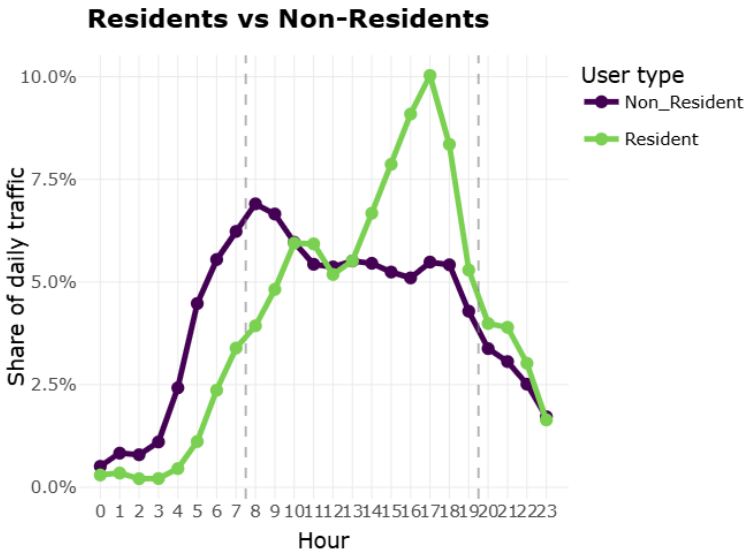
By comparing the three models it's clear that KNN must be excluded. Logistic Regression has shown the best model on average, and it will be used to build the resident identikit.

The confusion matrix for the Logistic Regression model confirms the difficulties of identifying the residents. Even using a balanced dataset, the model tends to classify a non-resident user as resident creating an overestimation effect.



The average resident" of Area C is statistically defined by the use of a private petrol-powered car and a preference for afternoon travel.

3.1.1.1. Time as predictor



Timestamps aren't the strongest variable to tell apart a resident from a non-resident. However, there is a clear distinction between the two categories that can be appreciated looking at the traffic trends of the two-user type. Non-residents mostly enter in the morning until 8:00, while residents return in the late afternoon around 17:00.

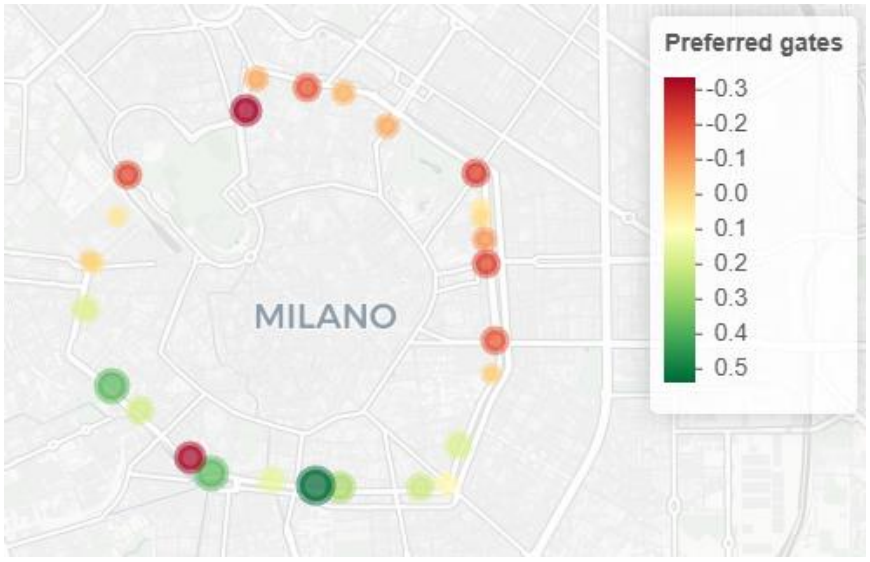
3.3.2. Preferred Gates

Research question	Are some gates preferred by the residents?
Dataset	Balanced 80/20 split dataset `df_features.rds`
Model	Lasso
File	`04-qualitative_data_analysis.rmd`

The final research module shifts the focus from general behavior to geographical specificity, identifying which of the 43 Area C gates are mostly patronized by residents versus non-residents

A Naive Bayes model was used to determine the probability of a user being a resident based on their choice of gate. To isolate the most significant gates from a large set of categorical predictors, a *Lasso*

Regression model was utilized since it's particularly effective selecting only the gates that are truly characteristic of residential behavior. Due to high computational requirements, the analysis was conducted on a balanced sample of 100,000 transits.



The Lasso model, combined with spatial mapping, reveals a clear geographical gap in gate usage. The south-west axis is the preferred one: highest coefficient values are concentrated in gates such as *Melegnano* and *Servio Tullio***.** These are located in historical residential districts characterized by roads not suitable for logistic purposes.

The east axis, on the opposite, show the lowest coefficient values. Major arteries like *Venezia* and *Monforte* are identified as primarily commercial routes, making them less appetible for resident entry.

Compared to the previously cloud gate approach, which has used *Naive Bayes*, the *Lasso* analysis has successfully fine-tuned the results, moving from general core gates to isolating specific, residential-only entry points.

The Lasso's performance reflects the difficulty of using only the gate as human behavior predictor: the model achieved a 41% accuracy rate. While seemingly low, this is considered a valid result for a multiclass analysis involving an unbalanced population of residents.

4. Conclusions

The analysis of Area C reveals a two-speed city: a Service Milan, dominated by polluting commuter and logistics flows that saturate the major access points (Venice, Turati), and a Residential Milan, which moves along lateral geographic routes. While pollution is a structural and predictable phenomenon linked to diesel/freight vehicles (LDA Accuracy: 79%), residential Milan is a spatial choice: the entrance point is the best predictor of concentration in the city center (Lasso Sensitivity: 74%). Area C functions as an environmental filter, but the real distinction between users is not the time of day, but the route chosen to access the city walls.

4.1. Answering with the data

Recalling the initial research questions, we're now able to decipher the Area C Traffic DNA.

4.1.1. Milan's habits

Does the rush hour correspond to office hours? Are there some anomalous peaks? The EDA analysis has confirmed that the traffic peaks match almost perfectly the rush hour with some night peaks due to logistic transport. Even if unpredictable and characterized also by external factors not considered in this analysis (such as weather), the traffic seems to be predictable and strongly correlated to the productive cycle.

What leads to more traffic? The crucial discriminant is the weekend: the traffic volume decreases a lot during the weekend, revealing that the Area C accesses are strongly related to the working patterns.

Which are clou-gates? Turati and Venezia are the main gates. With a mixed traffic typology, they are the gates used by the users. Due to their important traffic volume, it's difficult profile with accuracy, those gates are used by the deliveries, the tourists and the white collars.

4.1.2. Environmental aspects

Is the "Euro-rule" working? Is the most-pollutant-vehicles traffic decreasing? The traffic of pollutant vehicles is decreasing, making the Area C policies an effective contributor of pollutant vehicles reduction. The hard-to-die category, contributing the most to increase pollution, is the diesel one combined with the trucks one.

In what context do the more-pollutant-vehicles access the most? The most pollutant vehicles (mostly trucks) access Area C during the morning, making the hour predictor the main cause of pollutant accesses. Having a destination of use equal to goods increases the probability as well.

4.1.3. User profiling

What does distinguish a resident? Is it the timestamp? The gate? The vehicle type? What distinguishes a resident isn't just the hour but the geography. The Lasso model has demonstrated that the gate used is the best predictor. The fuel type is also good predictors: a resident will probably drive petrol. As mentioned, the time isn't a good predictor, but a resident will probably enter Area C in the afternoon.

Are some gates preferred by the residents? Residents will access Area C in the overly populated gates such as *Melegnano* and *Servio Tullio*. They usually avoid gates such as *Italia* and *Magenta*.