Developing a Decision Support System with a Georeferenced Smart City Security Index (SCSI): A Case Study of Messina

Giuseppe Accardo[1, *,†], Roberta Marino[1,*,†] and Valentina Esposito[1,*†]

[1] *Data Jam srl, Centro Direzionale Isola F8, Via F. Lauria, Naples, 80143, Italy*

Abstract

With the rapid growth of urban population, cities are facing increasing challenges in terms of mobility, sustainability, and living conditions. Smart cities leverage advanced technologies to improve urban efficiency and citizens' quality of life.

This work aims to empower the Public Administration (PA) of Messina, a medium-sized Italian city, with a georeferenced Smart City Security Index (SCSI) to monitor urban security and inform decision-making processes.

To achieve this, we trained a Random Forest Regressor using open data alongside territory specific key performance indicators (KPIs) and insecurity indicators. The model assigns a security score from 0 to 100 to each city area, achieving a Root Mean Squared Error (RMSE) of 5.6 on the test set.

Furthermore, integrating the model with a Decision Support System (DSS) allows PA members to assess changes in the SCSI in response to adjustments made to the input factors, supporting decision-making.

Keywords

smart city, open data, decision support system

1. Introduction

This work aims to leverage Artificial Intelligence (AI) to develop a specific smart city index for monitoring urban security in Messina, ultimately contributing to a smarter city.

The concept of a "smart city" encompasses the integration of technology and urban planning to enhance a city's sustainability, efficiency, and innovation. Several Smart City Indices (SCIs) have been developed in the literature to assess and quantify these aspects. These indices typically consider a range of services and projects that contribute to a city's "smartness," encompassing areas like public safety (e.g., reduced traffic accidents) and environmental sustainability.

SCIs function by aggregating multiple variables and indicators into a single score, providing a statistical summary of a city's overall performance. Monitoring this score over time allows for evaluation of a city's progress in achieving its "smart city" goals.

Table 1 summarizes some of the most widely recognized SCIs from the literature.

AI, on the other hand, has become a crucial tool for researchers in smart city initiatives. This, coupled with the open data movement, has spurred further research using these sophisticated techniques to unlock the potential of data in realizing smart city goals.

There is some evidence of positive impacts in the transportation, sustainability, or security fields [7][8][9][10][11][12].

This work aims to equip the Public Administration (PA) of Messina with a tool for monitoring urban security and informing decision-making processes. This tool leverages a georeferenced and machine learning-based Smart City Security Index (SCSI)

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy

[*] Corresponding author.

[†] These authors contributed equally.

gi.accardo@almaviva.it;
r.marino@almaviva.it;
v.esposito@almaviva.it

Table 1
Smart Cities Indexes in the literature.

Index	KPI
Arcadis Sustainable Cities Index [1]	20 indicators
Innovation Cities Index [2]	162 indicators
ISO 37120 [3]	100 indicators
ITU FG-SSC [4]	88 indicators
Networked Society City Index [5]	35 indicators
Siemens Green City Index [6]	30 indicators

2. Materials and Methods

This section details the data sources utilized for this study. We describe the steps involved in constructing the variables that will be employed by the machine learning (ML) model. Additionally, we present an overview of the exploratory analyses conducted to gain insights into the characteristics of the dataset.

The city is subdivided into 287 spatial units (tiles), each encompassing an area of 1 km^2. The SCSI will be used to assess the security level of each tile over time.

It follows that each feature within the dataset must adhere to a specific structure, consisting of a unique triad: geometry_id, month, and year. The year and month fields represent the reference time, while the geometry_id field uniquely identifies a tile.

2.1. Open data

We utilized open data from the city of Messina, which are described in the following section.

Municipal Police measures gather data on accidents involving traffic violations.

As an initial data preprocessing step, we addressed missing geospatial coordinates. We leveraged the Nominatim open-source API [13] to geocode these locations using the information provided in the "Luogo Incidente" (incident location) text column. Prior to geocoding, the text data underwent cleaning procedures using natural language processing (NLP) techniques. This process successfully assigned geographic coordinates to 84% of the previously unknown locations. Next, we extracted the variables of interest by aggregating the data by geometry_id, year and month based on the articles of traffic violation, according to the regulation in Italy.

This resulted in the following features:

- “prov_precedenza” (precedence) obtained as the sum of incidents with violations of articles 145 and 150.
- “prov_velocita” (speed) considers only the violation of Article 141.
- “prov_posizione” (position) obtained as the sum of articles 154, 149, 143, 148 and 144.
- “prov_documenti” (documents) as the sum of Articles 80, 193, 116, 180, 126, 94 and 93.
- “prov_sosta” (stop) derived as the sum of articles 158 and 157.
- “prov_segnaletica” (signals) derived as the sum of incidents with violation of Articles 40, 41 and 146.

Like the approach used for Municipal Police measures data, we addressed missing geospatial coordinates within the Lighting Points data. We employed the Nominatim open-source API for geocoding, using the information provided in the "Ubicazione toponomastica" (toponomastic location) text column. As with the previous data source, text cleaning procedures were necessary prior to geocoding, leveraging NLP techniques. This process successfully assigned geographic coordinates to 78% of the locations where coordinates were previously missing. Next, the feature of interest, namely the number of public lighting poles present in a certain time tile (“n_pali_luce”), was calculated by summing the poles falling by geospatial coordinates in the analyzed tile.

Urban Video surveillance details the closed-circuit television (CCTV) system operating within the Municipality. The data concern only administration-owned cameras, all of which are georeferenced, and have no missing values. Here, the variable of interest is the number of cameras present in a specific time tile (“n_telecamere”). We obtained this value by summing the CCTVs that fall within the analyzed tile, based on their geospatial coordinates.

2.2. Digital exhaust data

For the construction of the features, in addition to the open data, we derived the following geolocated indicators that can characterize tiles in the city of Messina.

The “sentiment” index is a measure of sentiment calculated on online content from the analysis period within the selected tile. It ranges from 0 to 100.

The “footfall” score is an absolute, and unlimited index that measures the foot traffic and popularity of a tile. This indicator considers various factors, such as the number of geolocated reviews, content on social media and aggregated and anonymized data originated from mobile devices.

The remaining features: "degrado" (degradation), "incendio" (arson), "incidente" (accident), and "crimini" (crimes), sum up the number of events linked to each of these categories per tile, year, and month. We collected this information by web-scraping from open and licensed/authorized closed sources such as websites blogs, social media and Police.

2.3. Data Preparation

After integrating the data described in the previous sections into a single table, we obtained a dataset with 12628 records, each representing a unique triad of geometry_id, month, and year.

The dataset refers to the time frame January 2019-August 2022, extremes included.

We then proceeded to analyze the content of this dataset, focusing initially on the target variable for the machine learning model, namely the "Security_Target".

This variable, is a weighted average of a qualitative and a quantitative index, representing the security level of each tile. The qualitative index considers the sentiment of online reviews related to security falling within each tile, while the quantitative index reflects the number of crimes committed. The qualitative index is weighted by the number of reviews in each tile, normalized between 0 and 1, while the quantitative index has a constant weight of 1. Values of the target variable range from 0 (lowest security) to 100 (highest security).

Figure 1 illustrates that for specific month and year, the target variable often takes the value of 100, which corresponds to the highest security level. Furthermore, as shown in Figure 2, the distribution of the target variable, considering the entire dataset, exhibits a significant imbalance, with the value 100 being the most frequent by a considerable margin.

To further explore the distribution of the target variable, we visualized it after excluding tiles with the highest security level (value 100). As shown in Figure 3, the remaining values exhibited a wider range, suggesting a more informative distribution for analysis.

Nevertheless, it was necessary to consider how to correct the imbalance in the values assumed by the target.

To understand the cause of this imbalance, we examined the features associated with tiles having the highest “Security_Target” (value 100). Interestingly, we discovered that 7812 records possessed identical features. In all these cases, the feature values were either 0 (indicating no events like for instance arson) or NaN (meaning data on factors like footfall and sentiment was unavailable). Due to these missing or non-informative features, we opted to remove these duplicate rows.

We obtained a dataset with 4816 records, 3398 of which were with target 100.

Following the initial data exploration, we analyzed the prevalence of missing values across all features (percentages shown in Table 2). To address this issue, we excluded observations where both sentiment and footfall data were missing. This exclusion step resulted in a dataset of 4654 records. Subsequently, the data was split into training and test sets. The training set comprised 3257 records, while the test set contained 1397 records.

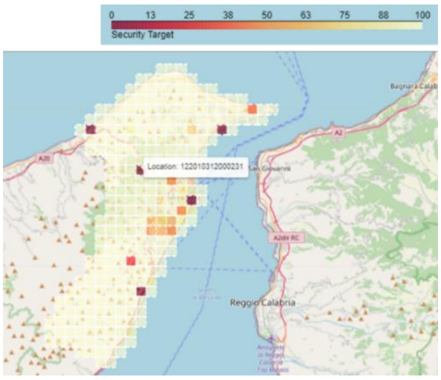


Figure 1: “Security_Target” distribution in Messina. This figure depicts the spatial distribution of the target. Color intensity is used to represent the “Security_Target” value, with light yellow indicating areas with the highest security level and dark red indicating areas with the lowest security level.

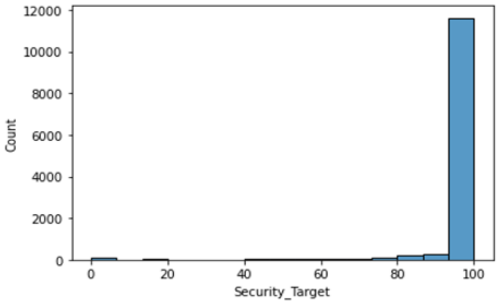


Figure 2: “Security_Target” Histogram.

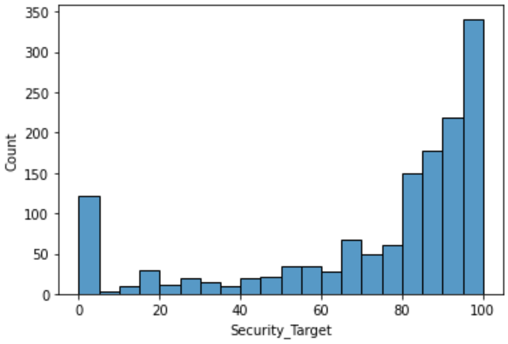


Figure 3: “Security_Target” with values less than 100. Histogram.

Table 2
Percentage of missing values

Feature	Percentage of missing values
prov_precedenza	0
prov_velocita	0
prov_posizione	0
prov_documenti	0
prov_sosta	0
prov_segnaletica	0
n_telecamere	0
n_pali_luce	0
sentiment	3.36
footfall	3.36
degrado	0
incendio	0
incidente	0
crimini	0

3. Results

This section details the ML model which was selected to compute the SCSI. This is a random forest regressor from the library scikit-learn, whose hyperparameters are indicated in Table 3. Analyzing the performance metrics of the ML model in Table 4, the residuals in the test set in Table 5 and the distribution of observed and predicted values in Figure 4 we assessed its goodness.
Having established the validity of the chosen model, we proceeded to analyze the impact of each feature on the target variable. Shapley Additive exPlanations (SHAP) values provide a useful graphical representation of these feature importances [14]. A beeswarm plot effectively visualizes the distribution of SHAP values, highlighting the features that exert the strongest influence on the model's predictions. Our analysis in Figure 5 reveals that the "degrado" feature has the greatest impact. High values of "degrado" (represented by red in the beeswarm plot) are associated with a lower SSCI, and vice versa. Similarly, the "n_pali_luce" feature is the second most important, with lower values corresponding to a reduced SSCI. This analysis of feature importance provides key insights into the behavior of the decision-support system (DSS). Following model development, we equipped the Public Administration of Messina with a DSS that enables them to simulate the impact of changes in the SSCI by modifying features within selected city tiles (see Figure 6 and Figure 7). In essence, these features function as controllable parameters that can be adjusted to improve the security level in specific areas.
Building on a similar approach, we developed a georeferenced green index (GI) for the PA of Messina (see equation (1)). This index assigns a score between 0 and 100, quantifying the overall quality and quantity of urban green space for each spatial unit. Similar to the SCSI, the green index is designed for integration with a DSS (see Figure 8 and Figure 9). However, unlike the SCSI, it does not employ machine learning techniques.
Below the expression to calculate the GI:

$$GI(tile) = \frac{w1 * UG + w2 * (\frac{HGA + TCA * \alpha}{ELA} * 100)}{w1 + w2} \qquad (1)$$

Explanation of variables:

1. UG (Urban green perception index): This index reflects the perceived quality and user experience of urban green spaces, derived from analyzing online reviews.
2. HGA (Horizontal green area, m^2): Represents the area of gardens, parks, and forests within the spatial unit.
3. TCA (Tree canopy area, m^2): Calculated as the sum of canopy area for all trees in the spatial unit.
4. ELA (Emerged land area, m^2): Represents the total land area excluding water bodies within the spatial unit.
5. α (Weight relative to the vegetative state of the canopy area): Derived from Visual Tree

Assessment (VTA) data. It is calculated as the weighted sum of the areas of tree crowns within a tile, adjusted for their vegetative state, divided by the total area of all tree crowns in the tile.

6. w1 and w2: Weights assigned such that the quantitative dimension (HGA and TCA) contributes twice as much as the qualitative dimension (UG) to the overall GI score.

Overall, this project demonstrates the value of data-driven approaches in urban planning. The SCSI and DSS empower the PA to make informed decisions regarding security, and the future integration of machine learning into the Green Index holds further promise for comprehensive urban management.

Table 3
Hyperparameters for the Random Forest Regressor

Hyperparameter	Value
n_estimators	100
oob_score	True
criterion	'squared_error'
max_depth	None
random_state	0
max_features	None
min_samples_split	6

Table 4
Performance metrics for the Random Forest Regressor, namely MAE (Mean Absolute Error), MSE (Mean Squared Error), and RMSE (Root Mean Squared Error). The Validation errors represent the mean of errors calculated during the 5-Fold cross-validation process.

Measure	Train	Validation (mean)	Test
MAE	1.01	2.08	1.78
MSE	9.76	40.11	31.09
RMSE	3.12	6.28	5.58

Table 5
Distribution of observed, predicted values and residuals considering data in the test set. Residuals are the difference between observed values and predicted values.

Value	observed	predicted	residual
count	1397	1397	1397
min	0	0	-40.99
25%	97.04	97.39	0
50%	100	99.95	0
75%	100	100	0.39
max	100	100	80.04

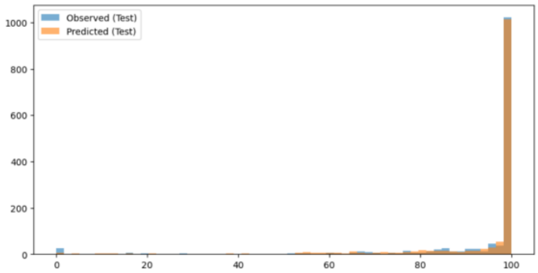


Figure 4: Distribution of observed and predicted values in the test set.

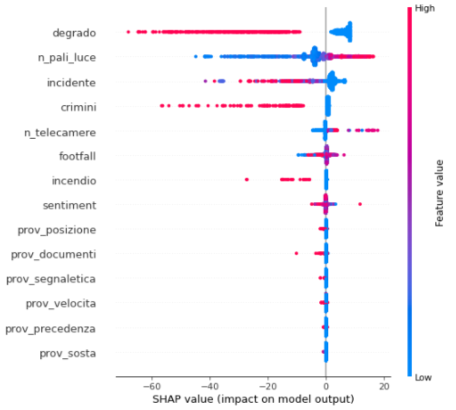
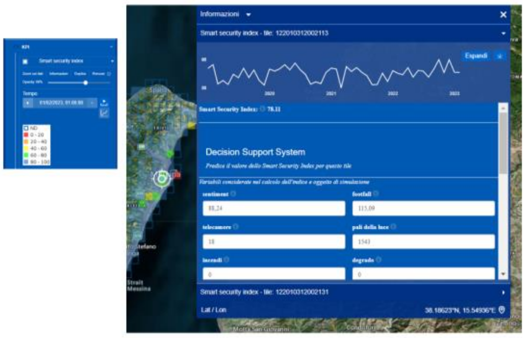


Figure 5: The "beeswarm" graph for the Random Forest regression related to the Smart Security City Index.



Smart security index - tile: 122010312002113

Smart Security Index: 78.11

Decision Support System

Predice il valore dello Smart Security Index per questo tile

Variabili considerate nel calcolo dell'indice e oggetto di simulazione

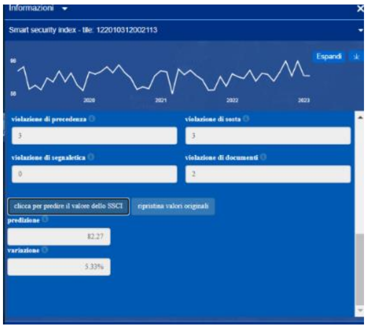
sentiment	footfall
88,24	115,09
telecamere	pali della luce
18	1543
incendi	degrado
0	0

Smart security index - tile: 122010312002131

Lat / Lon 38.18623°N, 15.54936°E

Figure 6: Example of an implementation of the SCSI in the Municipality of Messina. Empty tiles indicate

areas with missing data for footfall and sentiment and the remaining features equal to 0.



Smart security index - tile: 122010312002113

violazione di precedenza: 3

violazione di sosta: 3

violazione di segnaletica: 0

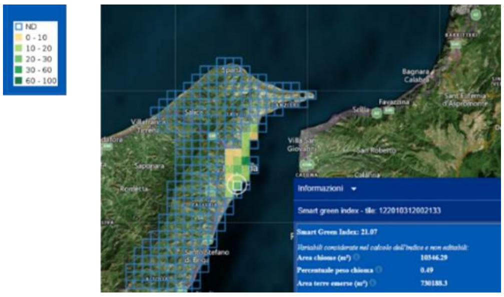
violazione di documenti: 2

clicca per predire il valore dello SSCI | ripristina valori originali

predizione: 82.27

variazione: 5.33%

Figure 7: Example of DSS application (security)



Smart green index - tile: 122010312002133

Smart Green Index: 21.07

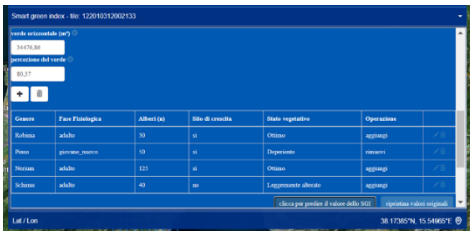
Variabili considerate nel calcolo dell'indice e non editabili:

Area chioma (m²): 10546.29

Percentuale peso chioma: 0.49

Area terre emerse (m²): 730188.3

Figure 8: Example of the GI implemented in the Municipality of Messina. Empty tiles represent areas with missing data for the municipal tree inventory. The number of tiles displayed will increase as the census continues.



Smart green index - tile: 122010312002133

verde orizzontale (m²): 34476,86

percezione del verde: 80,37

Genere	Fase Fisiologica	Alberi (n)	Sito di crescita	Stato vegetativo	Operazione
Robinia	adulto	30	si	Ottimo	aggiungi
Pinus	giovane_nuovo	50	si	Deperiente	rimuovi
Nerium	adulto	125	si	Ottimo	aggiungi
Schinus	adulto	40	no	Leggermente alterato	aggiungi

clicca per predire il valore dello SGI | ripristina valori originali

Lat / Lon 38.17385°N, 15.54965°E

Figure 9: Example of DSS application (urban green condition).

References

[1] Arcadis. (2022, June 21). The Arcadis Sustainable Cities Index 2022. [Member Spotlight]. Retrieved from https://www.arcadis.com/en/knowledge-hub/perspectives/global/sustainable-cities-index.

[2] 2thinknow. (2023). Innovation Cities™ Index. Retrieved from https://innovation-cities.com/worlds-most-innovative-cities-2022-2023-city-rankings/26453/

[3] International Organization for Standardization. (2018). ISO 37120:2018 Sustainable development of communities - Indicators for city services and quality of life.

[4] International Telecommunication Union (ITU). (n.d.). The Telecommunication Standardization Sector (ITU-T). Retrieved from https://www.itu.int/en/ITU-T/Pages/default.aspx

[5] Ericsson. (n.d.). Networked Society City Index. Retrieved from https://www.ericsson.com/en/reports-and-papers/networked-society-insights

[6] Siemens AG. (n.d.). Siemens Green City Index. Retrieved from https://assets.new.siemens.com/siemens/assets/api/uuid:cf26889b-3254-4dcb-bc50-fef7e99cb3c7/gci-report-summary.pdf

[7] Agarwal, P. K., Gurjar, J., Agarwal, A. K., & Birla, R. (2015). Application of artificial intelligence for development of intelligent transport system in smart cities. *Journal of Traffic and Transportation Engineering*, *1*(1), 20-30.

[8] Bharadiya, J. (2023). Artificial intelligence in transportation systems a critical review. *American Journal of Computing and Engineering*, *6*(1), 34-45.

[9] De Las Heras, A., Luque-Sendra, A., & Zamora-Polo, F. (2020). Machine learning technologies for sustainability in smart cities in the post-covid era. *Sustainability*, *12*(22), 9320.

[10] Hassan, S. I., & Agarwal, P. (2020). Analytical approach to sustainable smart city using IoT and machine learning. In *Big Data, IoT, and Machine Learning* (pp. 277-294). CRC Press.

[11] Lourenço, V., Mann, P., Guimaraes, A., Paes, A., & de Oliveira, D. (2018, July). Towards safer (smart) cities: Discovering urban crime patterns using logic-based relational machine learning. In *2018 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

[12] Butt, U. M., Letchmunan, S., Hassan, F. H., Ali, M., Baqir, A., Koh, T. W., & Sherazi, H. H. R. (2021). Spatio-temporal crime predictions by leveraging artificial intelligence for citizens security in smart cities. *IEEE Access*, *9*, 47516-47529.

[13] OpenStreetMap contributors, "Nominatim," OpenStreetMap wiki, 2023, https://nominatim.openstreetmap.org/.

[14] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 4768–4777.