Uncovering Key Influencers of Smart City Development Using Statistical Correlation and Association Rule Mining Techniques

Nasina Harshavardhan
Computer Science and Engineering
VIT-AP University
Amaravati 522241, Andhra Pradesh, India
harshavardhan.22bce9054@vitapstudent.ac.in

Yamini Kodali
Computer Science and Engineering
VIT-AP University
Amaravati 522241, Andhra Pradesh, India
yamini.24phd7012@vitap.ac.in

Y. V. Pavan Kumar, *SMIEEE*
Electronics Engineering
VIT-AP University
Amaravati 522241, Andhra Pradesh, India
pavankumar.yv@vitap.ac.in

***Abstract* - Aligned with the UN SDGs, there is a global initiative to transform conventional cities into smart cities for holistic national growth. This is a complex process influenced by various direct and/or indirect influencers, including mobility, environment, energy, education, governance, economy, culture, technology, and essential services for livability, viz., healthcare, housing, and security. Hence, this paper proposes a data-driven approach to systematically analyze the relationship between the key influencers and smart city development. This uses the Pearson correlation analysis for computing the correlation coefficients and the Apriori algorithm for association rule mining. A real-time dataset titled "Smart City Index Analysis (IoT)" from Kaggle is used for the proposed analysis. Simulation results reveal that the parameter "Smart_Living" exhibits the highest direct influence on the smart city development index ("SmartCity_Index"), with a 0.77 correlation coefficient. Also, the combinations of "Smart_Government" and "Smart_Living" exhibit the highest two-parameter indirect influence with a correlation coefficient of 0.33, while "Smart_Environment", "Smart_Government", and "Smart_Living" exhibit the highest three-parameter indirect influence with a correlation coefficient of 0.097. These findings help identify key influencers, thereby developing effective urban planning strategies for smart cities.**

Keywords - Association rule mining; Correlation analysis; Smart city; Smart city influencers; Statistical correlation; Urbanization.

I. Introduction

Rapid urbanization has intensified the demand for smart cities, which play a crucial role in enhancing the quality of life, stimulating economic growth, and advancing sustainability. Smart cities utilize technology and data to enhance service delivery, optimize infrastructure, and foster resilient urban environments. The smart city index data aims to increase the quality of living by analyzing aspects as mobility, health, safety, transparency, smart management, waste management, etc. Various data-driven works carried out in the literature are described as follows, which are useful in different aspects for the growth of smart cities.

A data-driven correlation analysis between city-level carbon emissions and power systems, setting the groundwork for environmental analytics in smart cities, was given in [1]. This also provides quantitative evidence of how smart city initiatives improve public services in China, establishing the benefits of urban digitalization [2]. A few of them explore context-awareness in smart cities through social media integration, emphasizing citizen-centric real-time insights [3]. To showcase these, a case study was applied on the impact of smart economy implementation on SME growth in Indonesia and Malaysia [4]. Further, some works highlight how smart economic models contribute to complete urban transformation [5], [6]. The aim of having all these statistical approaches is to reduce the waste, emissions, and develop green infrastructure in smart cities, which designs e-waste data models to support sustainable waste management [7]. It advocates ecological sustainability and waste management for AI-driven renewable energy systems in Europe's smart cities [8]. In smart cities and especially in smart homes, the smart energy systems and data-driven consumption analytics reveal various useful insights to consumers [9]. It uses ANOVA to study the variance in energy consumption data in smart homes and smart cities [10]. Further, it is also emphasized to detect redundancies in data to improve efficiency in smart cities [11]. Further, various statistical techniques are used on smart home data for stability assessment [12]. A Pearson correlation-enhanced machine learning model was applied in smart cities, especially in cyberattacks and IoT environments [13]. It develops AI-driven solutions for privacy and security in smart city databases [14]. It was also applied to improve video stabilization that is relevant to surveillance and monitoring of smart cities [15]. In medicine, this analysis is used for synchrony detection of EEG Signals, demonstrating statistical methods in smart city health systems [16].

Considering all the abovementioned developments and as per the current knowledge that is required to effectively manage the resources for the development of future smart cities, this paper proposes a correlation analysis based on the combination of Pearson correlation and Apriori algorithm to identify all direct and indirect influencers of smart city growth. The Pearson correlation method is simple, fast, widely used, requires minimal data, and involves fewer assumptions compared to partial correlation and regression analysis. Moreover, it is well-suited for preliminary analyses before progressing to more complex causal modeling techniques such as path analysis.

II. Dataset Description

To implement the proposed correlation analysis, a dataset named "Smart City Index Analysis (IoT)" from the Kaggle database is considered [17]. It is a CSV file, containing various columns, namely, City, Country, Smart_Economy, SmartCity_Index, Smart_Environment, Smart_Government, Smart_Mobility, Smart_Living, and Smart_People. This

describes the data collected around the world in 2020. Here, the City includes a list of smart cities across the world. The country column includes the list of countries where smart cities across the world are located. Smart_Mobility includes the index calculated from the assessment of the city-wide public transportation system, communication and information technology accessibility. Smart_Environment includes the index calculated from environmental sustainability impact, energy, and pollution management. Smart_Government includes the index calculated from the comparative study of transparent governance and open data initiatives of smart cities across the world. Smart_Economy includes the index calculated through global comparison of city-wide productivity, economic vitality. Smart_People includes the index calculated by comparing social and cultural plurality, the educational system, and its supporting ancillary services. Smart_Living includes the index calculated by measuring metrics around healthcare services, social security, and housing quality. SmartCity_Index includes the aggregate score for the smart city model based on smart city supergroups.

Out of all the columns of the dataset, the columns, namely, Smart_Mobility, Smart_Environment, Smart_Government, Smart_Economy, Smart_People, and Smart_Living, are considered as the influencing parameters on the target parameter SmartCity_Index, in this work.

III. Methodology

The proposed study involves a multi-level correlation analysis framework to identify both direct and indirect influencing parameters on the target parameter. The implementation flow is shown in Fig. 1. The process starts with data loading and data preprocessing. Initially, the DataFrames are created, and then the Pearson correlation analysis and Apriori algorithm-based association rule mining are performed. The Pearson correlation analysis quantifies the relation between any one influencing parameter and the target parameter by computing the Pearson correlation coefficient (PCC) [15]. The Apriori algorithm identifies the hidden relationships by finding the associations between all the influencing parameters and the target parameter. This algorithm finds the parameters which are occurring more commonly and associations among the parameters [12].

The parameter that directly influences the target parameter is assumed as the direct influencer, and all such parameters are identified and quantified at level-1 analysis. Similarly, the parameter that influences the target parameter via another parameter(s) (intermediate influencer(s)) is assumed as an indirect influencer. All such combinations of parameters that indirectly influence the target parameter are identified and quantified at levels 2 and 3. Here, level-2 indicates the combination of two influencing parameters, and level-3 indicates the combination of three influencing parameters that are considered at a time. While calculating the PCC values, the influencing parameters are automatically selected by filtering for numerical values, excluding the target parameter. All these computed PCC values are stored in a dictionary. The range of the PCC is from -1 to +1, where the values -1 to 0 specify negative correlation, 0 to 1 specify positive correlation, and 0 specify no correlation between the parameters. Further, it is to be noted that the highest positive coefficient value (closer to 1) indicates the strongest correlation (linear relationship), while the highest negative coefficient value (closer to -1) indicates the weakest correlation (nonlinear relationship), and values around 0 signify little to no linear relationship between the two parameters.

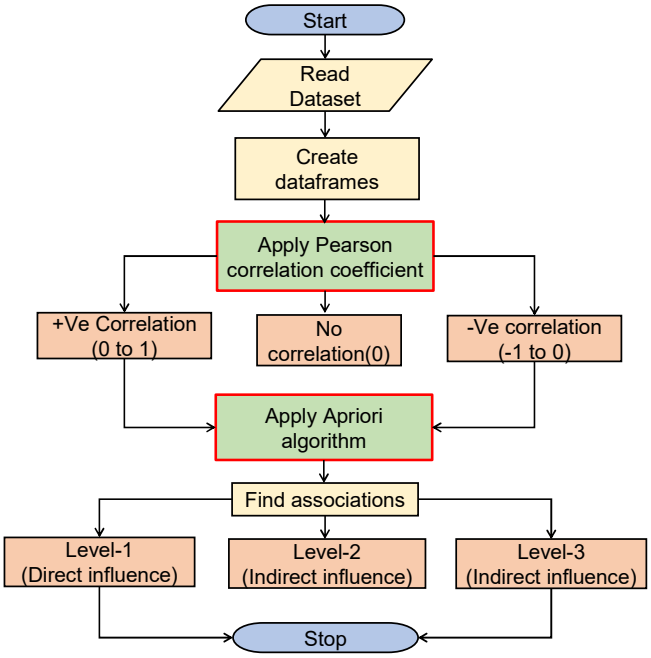

Fig. 1. Implementation flow of the proposed methodology.

Besides, it is to be noted that, in a positive correlation, both parameters (influencer and the target) tend to increase together, whereas in a negative correlation, an increase in one parameter is typically followed by a decrease in the other. Conversely, a decrease in one parameter often corresponds to an increase in the other, leading to a correlation coefficient value that may approach zero when the relationship is negligible. In the case of indirect influencers, the overall PCC value is computed as a product of the direct PCC values of each pair of parameters.

IV. Results and Discussion

The correlation network showing the connectivity of various influencers on the SmartCity_Index is shown in Fig. 2. In this figure, the green colour line shows the positive correlation between the connected parameters, and the red colour line shows the negative correlation between the connected parameters. For example, Smart_Mobility has a positive correlation with the parameter Smart_People, with a correlation coefficient of 0.35. Similarly, the parameter Smart_Mobility has a negative correlation with the parameter Smart_Economy, with a correlation coefficient of -0.26. Further, it is also noticed that the thick line indicates a strong influence, and the thin line indicates a weak influence between the parameters. For example, the parameter Smart_Living has the highest correlation with the parameter SmartCity_Index, represented with a thick line. Similarly, the parameter Smart_Economy has the lowest correlation with the parameter Smart_Living, which represents a thin line. The cumulative correlation matrix is given in Fig. 3, showing the relationships between multiple parameters relevant to smart city development. This matrix quantifies the strength between any two parameters using PCC values, as discussed in "Section III: Methodology". Moreover, the matrix allows for the scrutiny of pairwise relationships among all considered parameters.

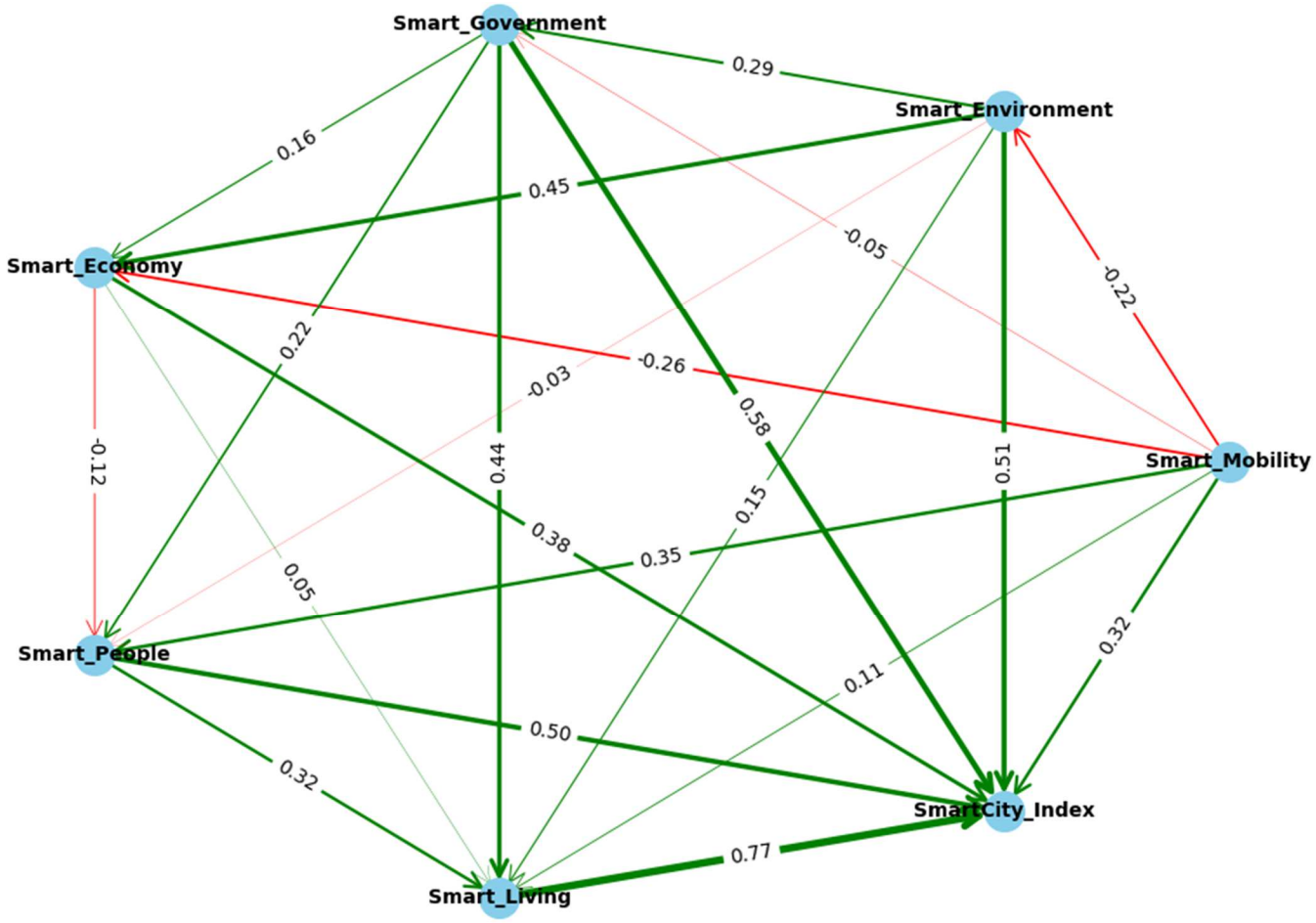

Fig. 2. Correlation network showing the connectivity of various influencers on the smart city index.

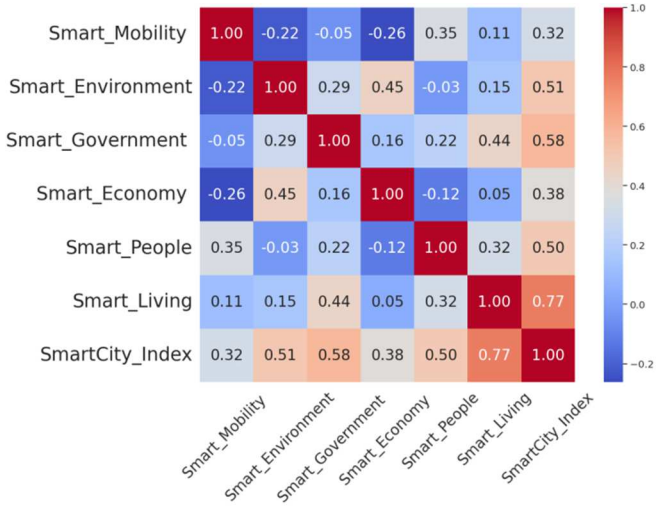

Fig. 3. Correlation matrix showing correlation between various parameters.

For instance, the correlation between Smart_Mobility and Smart_Environment is -0.22, denoting a weak negative correlation. As mentioned in Section III, this negative correlation indicates that an increase in Smart_Mobility is typically accompanied by a decrease in Smart_Environment or vice versa. This might point to potential trade-offs or conflicts between urban mobility solutions and environmental sustainability efforts. Similarly, the correlation between Smart_Mobility and Smart_Living is 0.11, denoting a weak positive correlation. Overall, the correlation matrix serves as a foundational tool in identifying both direct and indirect influencers of smart city components, guiding policymakers and urban planners in prioritizing interventions for balanced and sustainable urban development. Using Figures 2 and 3, an analysis of direct (Level-1) and indirect (Level-2 and 3) influencers of the SmartCity_Index is explained in the following subsections.

A. Analysis of Direct (Level-1) Influencers

The extent of direct influence exerted by various parameters on the SmartCity_Index is illustrated in Fig. 4.

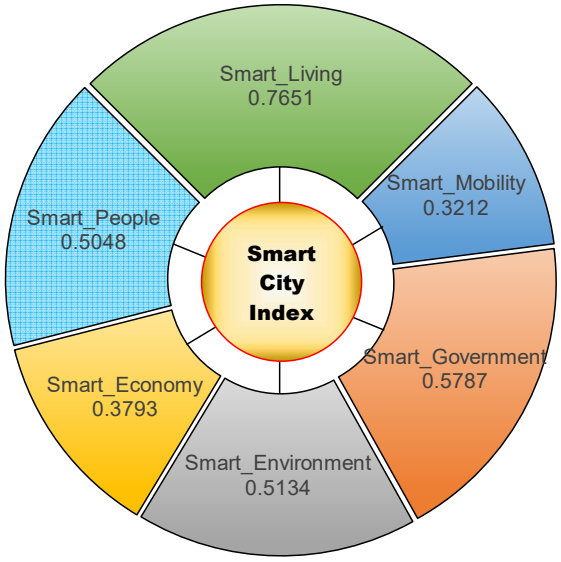

Fig. 4. Correlation of direct (level-1) influencers on the smart city index.

Among these direct influencers, Smart_Living exhibits the strongest positive correlation on the SmartCity_Index, with a PCC value of 0.77, indicating a significant impact. This implies that improvements in Smart_Living parameter (which includes factors such as healthcare, safety, housing, and quality of life) have a significant and direct positive impact on the overall smart city performance. Similarly, Smart_Mobility exhibits the weakest correlation on the SmartCity_Index, with a PCC value of 0.32. However, for this considered case study, it is observed that there is no influencer exists with a negative PCC value. This indicates that all influencing parameters have at least some direct influence with a positive PCC value on the target parameter, i.e., on SmartCity_Index.

B. Analysis of Indirect Influencers

The amount of indirect influence on SmartCity_Index by the combination of two parameters (level-2) and three parameters (level-3) is explained in this section.

1) Level-2 Influencers:

In Table I, the PCC values of level-2 are presented. The table contains both positive and negative PCC values. The strength of the indirect influencers on the target parameter is calculated as the multiplication of the two PCC values, which is in the range from -1 to +1. All level-2 indirect influencers and their PCC values are depicted in Fig. 5. From this figure, for example, a cumulative negative PCC value of -0.07 from Smart_Environment to SmartCity_Index is obtained as a multiplication of -0.22 (the PCC from Smart_Environment to Smart_Mobility) and 0.32 (the PCC from Smart_Mobility to SmartCity_Index). Similarly, a cumulative positive PCC value of 0.33 from Smart_Government to SmartCity_Index is obtained as a multiplication of 0.44 (the PCC from Smart_Government to Smart_Living) and 0.77 (the PCC from Smart_Living to SmartCity_Index).

Fig. 5. Correlation of indirect (level-2) influencers on the smart city index.

2) Level-3 Influencers:

A three-parameter combination analysis that explores indirect influences on the SmartCity_Index is illustrated in Fig. 6. This analysis investigates the interaction between two influencers via an intermediate parameter to assess their combined impact on smart city development. According to the findings presented, it is observed that the strongest indirect correlation is exhibited by the combination of the parameters Smart_Environment, Smart_Government, and Smart_Living on the target parameter SmartCity_Index, with a correlation coefficient of 0.097.

TABLE I. CORRELATION COEFFICIENT VALUES OF LEVEL-2 INDIRECT INFLUENCERS

Parameter1 (P1)	Parameter2 (P2)	Target Parameter (TP)	PCC (P1 to P2)	PCC (P2 to TP)	Cumulative PCC
Smart_Mobility	Smart_Government	SmartCity_Index	-0.04	0.57	-0.02
Smart_Mobility	Smart_Environment	SmartCity_Index	-0.22	0.51	-0.11
Smart_Mobility	Smart_Economy	SmartCity_Index	-0.26	0.37	-0.09
Smart_Mobility	Smart_People	SmartCity_Index	0.34	0.50	0.17
Smart_Mobility	Smart_Living	SmartCity_Index	0.11	0.76	0.08
Smart_Government	Smart_Mobility	SmartCity_Index	-0.04	0.32	-0.01
Smart_Government	Smart_Environment	SmartCity_Index	0.28	0.51	0.14
Smart_Government	Smart_Economy	SmartCity_Index	0.16	0.37	0.06
Smart_Government	Smart_People	SmartCity_Index	0.22	0.50	0.11
Smart_Government	Smart_Living	SmartCity_Index	0.44	0.77	0.33
Smart_Environment	Smart_Mobility	SmartCity_Index	-0.22	0.32	-0.07
Smart_Environment	Smart_Government	SmartCity_Index	0.28	0.57	0.16
Smart_Environment	Smart_Economy	SmartCity_Index	0.45	0.37	0.17
Smart_Environment	Smart_People	SmartCity_Index	-0.03	0.50	-0.01
Smart_Environment	Smart_Living	SmartCity_Index	0.15	0.76	0.11
Smart_Economy	Smart_Mobility	SmartCity_Index	-0.26	0.32	-0.08
Smart_Economy	Smart_Government	SmartCity_Index	0.16	0.57	0.09
Smart_Economy	Smart_Environment	SmartCity_Index	0.45	0.51	0.23
Smart_Economy	Smart_People	SmartCity_Index	-0.12	0.50	-0.06
Smart_Economy	Smart_Living	SmartCity_Index	0.04	0.77	0.03
Smart_People	Smart_Mobility	SmartCity_Index	0.34	0.32	0.11
Smart_People	Smart_Government	SmartCity_Index	0.22	0.57	0.12
Smart_People	Smart_Environment	SmartCity_Index	-0.03	0.51	-0.01
Smart_People	Smart_Economy	SmartCity_Index	-0.12	0.37	-0.04
Smart_People	Smart_Living	SmartCity_Index	0.32	0.77	0.24
Smart_Living	Smart_Mobility	SmartCity_Index	0.11	0.32	0.03
Smart_Living	Smart_Government	SmartCity_Index	0.44	0.57	0.25
Smart_Living	Smart_Environment	SmartCity_Index	0.15	0.51	0.07
Smart_Living	Smart_Economy	SmartCity_Index	0.04	0.37	0.01
Smart_Living	Smart_People	SmartCity_Index	0.32	0.50	0.16

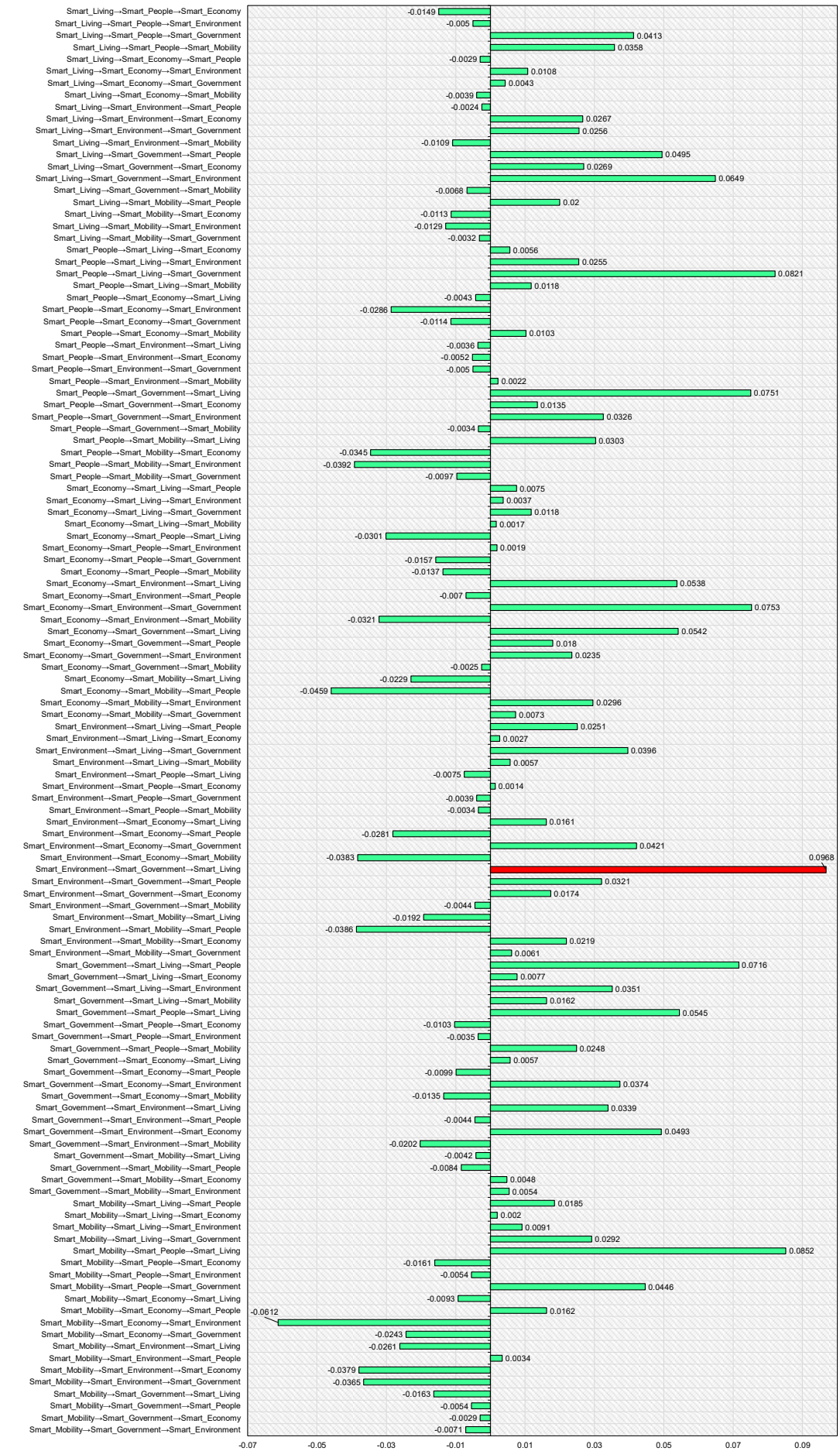

Fig. 6. Correlation of indirect (level-3) influencers on the smart city index.

This indicates a weak positive relationship, suggesting that improvements in environmental initiatives may indirectly support better living standards when mediated by effective governance. On the other hand, the weakest indirect correlation is found by the combination Smart_Mobility, Smart_Economy, and Smart_Environment on the target parameter SmartCity_Index, with a correlation coefficient of -0.06. This suggests a weak negative indirect relationship, possibly indicating misalignment or inefficiencies between transportation systems and environmental efforts when influenced by economic factors. Overall, these results underscore the complex and interconnected nature of smart city dimensions. While some indirect pathways show the potential for positive reinforcement, others reveal areas where strategic interventions are necessary to ensure alignment between governance, economy, environment, mobility, and living standards. Further, the correlation coefficients for all three-parameter combinations remain within a narrow range, from -0.06 to 0.097, indicating that the level of indirect influence on the SmartCity_Index through three-parameter combinations is minimal. Based on these observations, it is concluded that extending the correlation analysis beyond level-2 (i.e., involving more than two parameters of indirect influence) is unlikely to yield significant insights. Therefore, for practical purposes and interpretability, only level-2 indirect influences are considered significant.

V. Conclusions

This paper implements a systematic correlation analysis to identify the key direct and indirect influencers of smart cities, which helps to upgrade the smart cities by satisfying SDG-3. There are both positive correlations and negative correlations observed in this analysis. The salient observations made from the simulation results are summarized as follows.

- The parameter Smart_Living has the highest direct influence on the target parameter SmartCity_Index with a correlation coefficient value of 0.77.
- The parameter Smart_Government has the highest indirect influence on the target parameter SmartCity_Index, with a cumulative correlation coefficient value of 0.33 via the intermediate parameter Smart_Living.
- The parameter Smart_Environment has shown the highest indirect influence on the target SmartCity_Index with a cumulative correlation coefficient value of 0.097 via the intermediate parameters, namely Smart_Government and Smart_Living.

These findings offer a holistic framework that emphasizes both direct and indirect influences, providing valuable insights for policymakers and urban planners to guide strategic planning and sustainable smart city growth.

References

[1] Y. Gao *et al.*, "A Data-driven Correlation Analysis Method for Cities Carbon Emissions with Power System," in *2021 Smart City Challenges & Outcomes for Urban Transformation (SCOUT)*, Bhubaneswar, India: IEEE, Dec. 2021, pp. 209–213. doi: https://doi.org/10.1109/SCOUT54618.2021.00051.

[2] M. Zhou, H. Liu, and Z. Wang, "Can Smart City Construction Promote the Level of Public Services? Quantitative Evidence From China," *IEEE Access*, vol. 10, pp. 120923–120935, 2022, doi: https://doi.org/10.1109/ACCESS.2022.3221759.

[3] F. Purnomo, Y. Heryadi, F. L. Gaol, and M. Y. Ricky, "Smart city's context awareness using social media," in *2016 International Conference on ICT For Smart Society (ICISS)*, Surabaya, Indonesia: IEEE, Jul. 2016, pp. 119–123. doi: https://doi.org/10.1109/ICTSS.2016.7792860.

[4] I. D. C. Arifah, A. Safitri, H. Fazlurrahman, G. K. Chen, A. N. Masrom, and F. Kharisma, "Smart Economy Implementation in Supporting SMEs Growth: Case Study in Indonesia & Malaysia Smart Cities," in *2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indonesia: IEEE, Dec. 2022, pp. 787–792. doi: https://doi.org/10.1109/ISRITI56927.2022.10053000.

[5] A. Youssef and P. Hajek, "The Role of Smart Economy in Developing Smart Cities," in *2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC)*, Rome, Italy: IEEE, Nov. 2021, pp. 276–279. doi: https://doi.org/10.1109/ISCSIC54682.2021.00057.

[6] S. Akaraci, M. A. Usman, M. R. Usman, and D. J. Ahn, "From smart to smarter cities: Bridging the dimensions of technology and urban planning," *2016 International Conference on Smart Green Technology in Electrical and Information Systems (ICSGTEIS)*, Denpasar, Indonesia, 2016, pp. 74-78, doi: https://doi.org/110.1109/ICSGTEIS.2016.7885770.

[7] M. P. De Novais, V. De Almeida Xavier, L. H. Xavier, and J. Hwang, "Modelling e-waste management data in smart cities," in *2021 2nd Sustainable Cities Latin America Conference (SCLA)*, Medellin, Colombia: IEEE, Aug. 2021, pp. 1–4. doi: https://doi.org/10.1109/SCLA53004.2021.9540096.

[8] A. C. Serban and M. D. Lytras, "Artificial Intelligence for Smart Renewable Energy Sector in Europe—Smart Energy Infrastructures for Next Generation Smart Cities," *IEEE Access*, vol. 8, pp. 77364–77377, 2020, doi: https://doi.org/10.1109/ACCESS.2020.2990123.

[9] K. Purna Prakash *et al.*, "A comprehensive analytical exploration and customer behaviour analysis of smart home energy consumption data with a practical case study," *Energy Reports*, vol. 8, pp. 9081–9093, Nov. 2022, doi: https://doi.org/10.1016/j.egyr.2022.07.043.

[10] Y. Kodali and Y. V. P. Kumar, "ANOVA-Based Variance Analysis in Smart Home Energy Consumption Data Using a Case Study of Darmstadt Smart City, Germany," *Engineering Proceedings*, vol. 82, no. 1, p. 31, Nov. 2024. doi: https://doi.org/10.3390/ecsa-11-20354.

[11] P. P. Kasaraneni, V. P. K. Yellapragada, G. L. K. Moganti, and A. Flah, "Analytical Enumeration of Redundant Data Anomalies in Energy Consumption Readings of Smart Buildings with a Case Study of Darmstadt Smart City in Germany," *Sustainability*, vol. 14, no. 17, p. 10842, Aug. 2022, doi: https://doi.org/10.3390/su141710842.

[12] P. P. Kasaraneni, Y. V. P. Kumar, and R. Kannan, "Data-driven analytics for power system stability assessment," in *Intelligent Data-Driven Modelling and Optimization in Power and Energy Applications*, 1st ed., Boca Raton: CRC Press, 2024, pp. 39–71. doi: https://doi.org/10.1201/9781003470274-3.

[13] A. Senthilkumar, S. Joshika, L. Santhi, S. K.S., and P. Charanarur, "Pearson Correlation Coefficient based Improved Least Square - Support Vector Machine for Cyber-Attack Detection in Internet of Things," in *2024 3rd Int. Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, Ballari, India: IEEE, Apr. 2024, pp. 1–4. doi: https://doi.org/10.1109/ICDCECE60827.2024.10549411.

[14] R. Jena, M. Kumar, I. R. Mallela, A. Das, S. Vashishtha, and N. Agarwal, "Artificial Intelligence-Driven Solutions for Privacy and Security in Smart City Data Management Systems for Smart Cities," in *2024 13th International Conference on System Modeling & Advancement in Research Trends (SMART)*, Moradabad, India: IEEE, Dec. 2024, pp. 197–203. doi: https://doi.org/10.1109/SMART63812.2024.10882164.

[15] J. Liu, Y. Zhang, and Q. Zhao, "Video stabilization algorithm based on Pearson correlation coefficient," in *2019 International Conference on Advanced Mechatronic Systems (ICAMechS)*, Kusatsu, Shiga, Japan: IEEE, Aug. 2019, pp. 289-293. doi: https://doi.org/10.1109/ICAMechS.2019.8861649.

[16] T. Zhou, Z. Mei, X. Zhu, and Z. Huang, "Synchrony Detection of Epileptic EEG Signals Based on Attention and Pearson's Correlation Coefficient," in *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Chengdu, China: IEEE, Oct. 2020, pp. 531-535. doi: https://doi.org/10.1109/CISP-BMEI51763.2020.9263624.

[17] https://www.kaggle.com/datasets/pythonafroz/smart-home-dataset, "Smart home dataset", Kaggle, 2020.