Artificial Intelligence Research and Development
A. Cortés et al. (Eds.)

doi:10.3233/FAIA220319

Deep Air – A Smart City AI Synthetic Data Digital Twin Solving the Scalability Data Problems

Esteve ALMIRALL [a,1], Davide CALLEGARO [a], Peter BRUINS [a],
Mar SANTAMARÍA [b], Pablo MARTÍNEZ [b] and Ulises CORTÉS [c]

[a] *Esade Business School, URL University*
[b] *300.000km, 300000kms.net*
[c] *Knowledge Engineering and Machine Learning Group, Universitat Politècnica de Catalunya*

Abstract. Cities are becoming data-driven, re-engineering their processes to adapt to dynamically changing needs. A.I. brings new capabilities, effectively enlarging the space of policy interventions that can be explored and applied. Therefore, new tools are needed to augment our capacity to traverse this space and find adequate policy interventions. Digital twins are revealing themselves as powerful tools for policy experimentation and exploration, allowing faster and more complete explorations while avoiding costly interventions. However, they face some problems, among them data availability and model scalability. We introduce a digital twin framework based on an A.I. and a synthetic data model on NO_2 pollution as a proof-of-concept, showing that this approach is feasible for policy evaluation and (autonomous) intervention and solves the problems of data scarcity and model scalability while enabling city level Open Innovation.

Keywords. Digital Twins, Smart City Policy, Synthetic data, Digital twins and synthetic data

1. Introduction

Digital Twins have been evolving in aerospace exploration, manufacturing, and many other areas and, together with them, has been an evolution of their understanding.

Two challenges that Digital Twins find in cities are the lack of complete data, particularly real-time data, and the need for scalability. Cities are large, grow and change constantly and have fuzzy borders. The idea of sensorizing a whole city is undoubtedly bold, difficult to attain, challenging to make it economically sound, and even more to keep it updated.

In this paper, we introduce "Deep Air," a proof-of-concept prototype to provide some light on solving these two problems using machine learning and synthetic data. In synthesis, our prototype is a digital twin for city pollution built with synthetic data cre-

[1] Corresponding Author: Esteve Almirall, Esade Business School, URL University; E-mail: esteve.almirall@esade.edu

ated by a calibrated machine learning model. We show that the accuracy of the prototype is enough for investigating pollution city policies and reactions to specific conditions, taking into consideration the low granularity of applicable procedures [1].

This approach will not only help solve some of the problems of Smart City digital twins but also enables a model of decentralized self-regulated governance based on AI and synthetic data digital twins that we believe are endowed with higher agility, flexibility, scalability, and far lower cost while enabling Open, data-driven innovation in cities.

2. Materials and Methods

A digital twin is a virtual representation of the characteristics and behaviors of a physical entity used to study and predict its conduct without having to experiment with the actual object [2].

Smart Cities is probably today one of the most substantial areas of development of digital twins, not only in terms of projects being developed across the world in cities.

Smart City digital twins heavily depend on an abundance of data, particularly real-time data, with fine granularity.

The central idea of our proposal is to use synthetic data. Therefore, data is created through an intermediary A.I.-based model to feed the digital twin together with real data from sensors and synthetic data (see Fig 1).

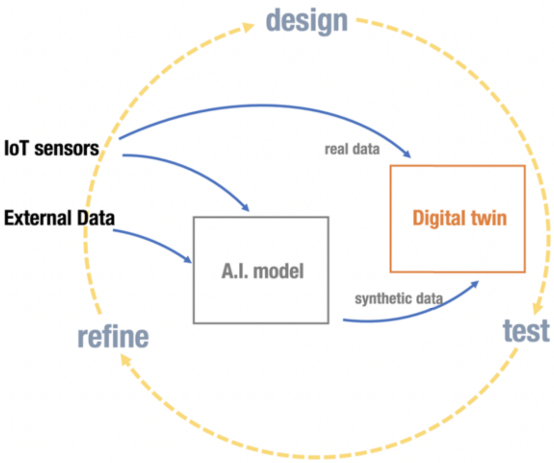

Figure 1. A digital twin fed with real and synthetic data in a design-test-refine loop.

The use of synthetic data will solve the problem of the scarcity of data. However, it could create another issue, the one of accuracy. We argue, however, that if the accuracy of the AI model supporting the digital twin is high enough and the granularity of the planning decision to be taken based on the data of the digital twin low enough, then there is a space where it will be no difference in terms of decisions between synthetic and real data. To express it formally, the equality of decision proposition must be satisfied.

Proposition 1. Equality of decisions.

Given a set of real data (measurements) $\Phi = \{\phi_1, \phi_2, \phi_3 \ldots\}$ and a set of synthetic data (generated measurements) $\hat{\Phi} = \{\hat{\phi_1}, \hat{\phi_2}, \hat{\phi_3} \ldots\}$ and given a digital twin $T(\Phi)$, where $T(\cdot)$ is a function assuming $\mathbb{R}^\infty$ values, we can define the enabling decision function as $D(T)$ over a discrete space of decisions $D(T) = \{\delta_1, \delta_2, \ldots\}$.

Thus, the equality of decisions proposition is satisfied when $D(T(\Phi)) = D(T(\hat{\Phi}))$, and this equality holds if and only if two conditions are met:

1. $\Phi \cong \hat{\Phi}$, therefore, a highly accurate model producing synthetic data is needed
2. $D = \{\delta_1, \delta_2, \delta_3, \ldots\}$ met when δ_i is highly separated from δ_j, $\forall_{i,j}$

In this paper, we will concentrate on the first condition 1), providing a proof-of-concept and showing that high accuracy (more than 88%) is possible in pretty complicated models (NO_2 congestion) with a minimal set of variables (eight in this case).

3. Results

Our results show that it is possible to predict NO_2 pollution data with an accuracy of 88.876%, std of 1.3768, with an XGBoost model primarily based on geographical data and only eight features. These are certainly encouraging results.

For this project, we have extracted data from multiple sources. Some data sets were publicly available, and others were given upon request.

- INE (Instituto Nacional de Estadística) data is all publicly available and found on their website [3].
- EEA (European Environmental Agency) data is publicly available and found on their website [4].
- The datasets held by 300000 km/s (300000kms.net) - an urban think tank located in Barcelona, which has been part of the research by providing insights and ad-hoc data - are private but not confidential and available upon request.

Initially, dataset collection resulted in approximately one hundred features. Using domain knowledge, an initial selection was made, resulting in a set of 28 elements that constituted our baseline model (Fig. 2).

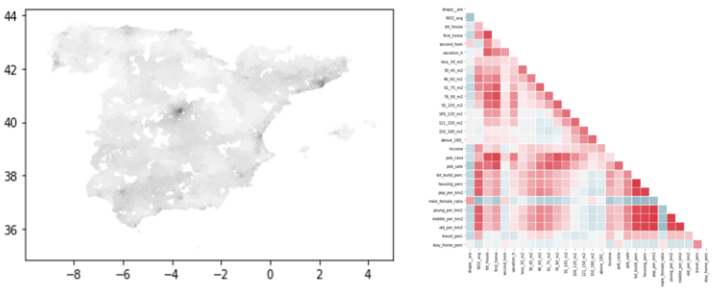

Figure 2. NO_2 data per district, together with the correlation matrix of the initial 28 features

Looking at feature importance, this model shows clear possibilities of simplification in addition to improvement using a more sophisticated algorithm.

We performed feature selection among the mix of lagged and non-lagged features, reducing them to only eight and using a Random Forest model.

Finally, we reproduce the same analysis with a more powerful model, an XGBoost, obtaining some improvement and reaching our final accuracy of 88.87%, std 1.376834 (See figure 3).

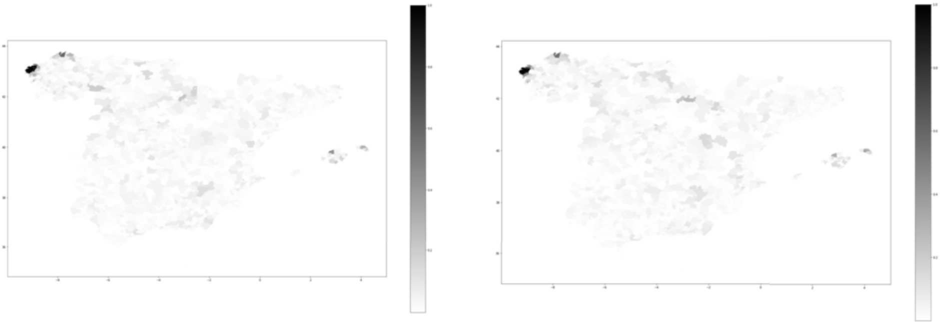

Figure 3. Random Forest and XGBoost model accuracy – darker is less accurate

As shown, only a few elements are determinants of NO_2 pollution, many of them static because they are part of the urban fabric. Therefore, it is possible to create models with limited features and high accuracy.

4. Conclusion

Through this paper, we have shown how synthetic data is feasible and solves many problems that today's digital twins face in complex socio-economic environments such as cities. Digital twins, in their broad sense as digital models that could accurately represent certain aspects of a city, allowing for experimentation, planning, knowledge discovery - metadata will be a valuable asset- and even near real-time adjustments or emergency activation procedures, are part of the future of management, optimization and city planning. Metaphorically we can say that code is the new concrete [5]. The proof of concept present in this work shows an alternative path that mixes real and synthetic data. We hope this research inspires a new generation of digital twins to support cheaper, more scalable, and open city management enabling open, data-driven innovation in cities.

References

[1] Almirall, E and Callegaro, D and Bruins, P and Santamaria, M and Martinez, P and Cortés, U uh. Deep Air. A Smart City AI Synthetic data Digital Twin cracking the scalability data problems; 2022.

[2] Jones D, Snider C, Nassehi A, Yon J, Hicks B. Characterising the Digital Twin: A systematic literature review. CIRP Journal of Manufacturing Science and Technology. 2020;29:36-52.

[3] Instituto Nacional de Estadistica. Census Data 2011; 2011. Available from: https://www.ine.es/censos2011_datos/cen11_datos_resultados_seccen.htm.

[4] European Environmental Agency. Air Pollution Data – EEA; 2019. Available from: https://www.eea.europa.eu/themes/air/dc.

[5] LADOT. Los Angeles. Technology Action Plan; 2019. Available from: https://ladot.io/wp-content/uploads/2019/03/LADOT-TAP-v7-1.pdf.