Towards a Cognitive Digital Twin of a Country with Emergency, Hydrological, and Meteorological Data

Jan Šturm
Jožef Stefan Institute
Jožef Stefan Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia
jan.sturm@ijs.si

Maja Škrjanc
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia
maja.skrjanc@ijs.si

Luka Stopar
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia
luka.stopar@ijs.si

Domen Volčjak
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia
domen.volcjak@gmail.com

Dunja Mladenić
Jožef Stefan Institute
Jožef Stefan Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia
dunja.mladenic@ijs.si

Marko Grobelnik
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia
marko.grobelnik@ijs.si

ABSTRACT

The paper presents a methodology for building a cognitive digital twin of a country elaborating on the conceptual design of a cognitive digital twin of a country. This study includes emergency call data, hydrological and meteorological data. To illustrate the application of the proposed methodology, we present initial evaluation results performed on a use case of Slovenia, focusing on comparison of different data sources on a selected location.

KEYWORDS

Cognitive Digital Twin, Real Time Data

1 INTRODUCTION

A cognitive a digital twin of a country is a digital model that replicates a nation's physical and social characteristics to simulate and forecast its behavior in diverse circumstances, utilizing historical data and real-time information. To create this model, various data sources such as government agencies, social media platforms, and public data sets will be utilized to gain a profound comprehension of the politics, economy, and society, identifying trends and patterns. Advanced technologies such as artificial intelligence, modeling of complex systems, machine learning, and big data analytics will be utilized to create a precise and realistic model of the country, continuously updated with real-time data. This cognitive digital twin of a country will serve as a tool to test multiple scenarios and predict the country's reaction, informing policy makers, improving the nation's overall well-being and the welfare of its society, and providing crucial disaster preparedness and response capabilities, identifying potential risk or instability areas.

2 RELATED WORK

The concept of a cognitive digital twin for a nation finds its roots in the broader realm of digital twin technologies, which traditionally pertained to replicating physical systems for simulation and predictive purposes. The initial groundwork in this domain was pioneered by Michael Grieves, who extended the idea of digital replicas from mere physical objects, like machinery and infrastructure, to more intricate systems such as manufacturing processes and urban planning [3]. Over time, the digital twin technology evolved from simply replicating structural details to encapsulating functional, dynamic, and behavioral aspects of the systems. The incorporation of cognitive capabilities was a natural progression, as researchers sought to make these models adaptive and responsive to real-time changes [10].

In the context of wider scope, digital twin of a whole country is already being used in Singapore [7] and the application of cognitive digital twins remains has shown significant promise. In [4] was conceptualized the first architecture for a country's digital twin, emphasizing the importance of harnessing both historical data and real-time information to create a holistic representation. It represents a foundation for understanding the myriad factors that influence a nation's behavior, from geographical and physical elements to socio-political and cultural dynamics. Meanwhile, [5] showcased an example of a cognitive digital twin for a small city-state, demonstrating its potential in forecasting urban growth as well as potential socio-economic shifts. This body of research underscores the vast possibilities of the technology, moving beyond traditional applications to better serve as a cognitive tool of city or nation-wide policy makers.

3 METHODOLOGY

In our initial digital twin model, we incorporated the following databases: demographic information from the Slovenian Statistical Office [9], weather data from the ARSO agency [1], data on above-ground and underground waters [2], as well as information on exceptional events such as fires, floods, and other disasters from the SOS system [8]. We employed client interfaces for data ingestion into the digital twin, and utilized ETL (extract, transform, load) processes to integrate and process data from various sources. Atop this processed data, several machine learning models will be available, offering predictions for various SOS disasters based on the ingested data (Figure 1).

3.1 Data Clients

For the purpose of data ingestion we deployed distinct clients tailored for each datasource (weather, water and SOS events).

Information Society 2023, 10–14 October 2023, Ljubljana, Slovenia

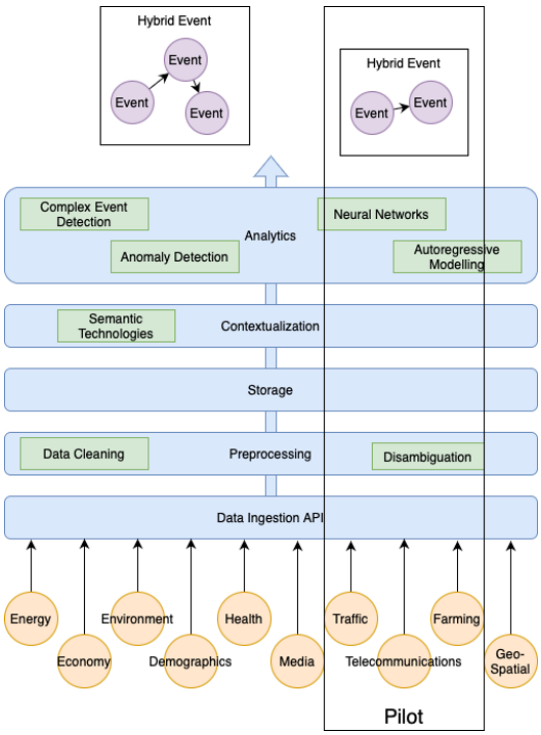

Figure 1: Conceptual design of cognitive digital twin of a country

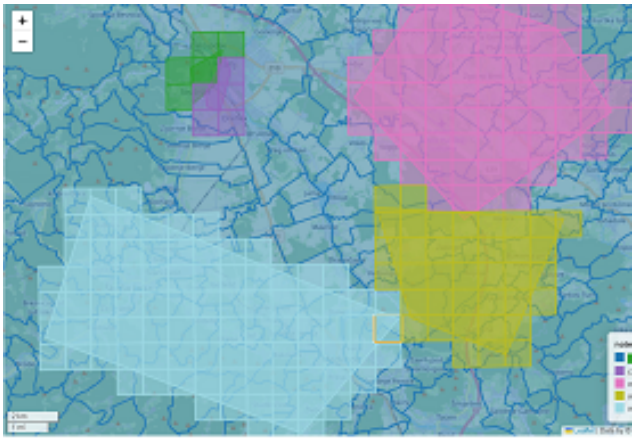

Figure 2: Conversion of geospatial formations into 1km x 1km squares

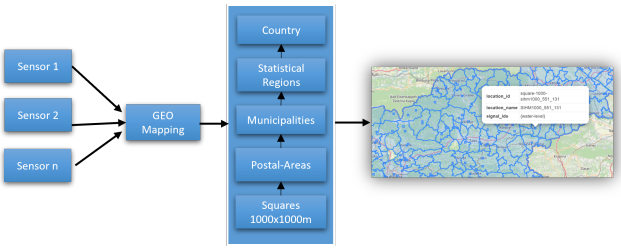

Figure 3: Spatial hierarchy

Each of these clients has a two-fold role. First, it fetches the raw data and channels it into the system. Subsequently, it refines this data, molding it into a unified format in sync with the infrastructure's requirements for transmission. Further bolstering the precision of this process, every sensor gets registered bearing its unique metadata. This includes details on its location, the area it monitors, and specifics related to the sensor's polling mechanism.

3.2 ETL Pipeline

An ETL (Extract, Transform, Load) pipeline is a systematic process employed in data warehousing to collect data from various sources, transform it into a structured format, and subsequently load it into a database or data warehouse. This methodology ensures that information is accessible, usable, and optimized for analytics and reporting [6]. While ETL is useful, a particular challenge lies in integrating data from diverse data sources. Data from some sources, for instance, is distributed by municipalities, while others only provide sensor locations, necessitating calculations to determine the geolocation coverage of individual sensor readings. Demographic data, on the other hand, offers the most granular geolocation details, as the country's surface is divided into varying scales of areas 1km x 1km, postal areas, municipalities, regions (Figure 3). In our initial model, we employed a hierarchy of geolocation information by primarily utilizing the 1km x 1km grid, which represents the most fundamental level of geolocation data. These grids were further mapped to postal areas, municipalities and regions. Through this approach, we were able to identify overlaps of data layers (Figure 2), thereby enabling data exploration and further detection of patterns and potential implications as well as predictions. Each layer represents a separate data source, which may contain information regarding population density, classifications of rural areas, and sensor readings.

3.3 Feature Engineering

Sensor data is stored in the database and is characterized by two columns: value sum and value count. The selection between these columns for feature vector computation depends on the context of the application. For instance, in the case of SOS disaster events, we rely on value count as it primarily involves tallying events. Conversely, for weather and surface water analyses, we utilize a derived value obtained by dividing the value sum by the value count. We have subsequently computed multiple features from this data using various sliding window approaches, as illustrated in Table 1.

4 EXPERIMENT

4.1 Dataset

Dataset in experiments includes SOS disasters, weather and surface water data, while other layers were not included in this paper. Data spans from January 1, 2010, to August 23, 2023. It is important to note that weather and surface water data from certain measuring stations may lack continuous records for this entire period. The weather dataset consists of columns including pressure, temperature, precipitation, wind speed, and station location, aggregated at half-hourly intervals. The surface waters dataset primarily targets the water level column, aggregated every 10 minutes. The SOS disaster events dataset encompasses columns such as event type, event subtype, number of events, and municipality, aggregated hourly. Data preprocessing encompasses two principal phases. Initially, data is categorized based on

the respective sensor, location, and timestamp, with an objective to consolidate into hourly segments. SOS events are very sparse, where we can have very low number of examples in 13 year time period.

4.2 Implementation Details

Experiments utilized Python 3.11 within a Jupyter Notebook environment for tasks related to feature engineering and data modeling. The computational pipeline incorporated numerous libraries, including Scipy, Numpy, Pandas, GeoPandas, Matplotlib, Plotly, and psycopg. Geospatial data, imported via psycopg, was seamlessly converted into a dataframe.

4.3 Experimental Results

The table 1 presents highest correlations associated with windbreaks in Ajdovščina. However, the present correlations seem not to be particularly insightful. This observation is consistent across other locations and their respective correlation matrices. A thorough refinement and meticulous preparation of the dataset, along with its associated features, would be indispensable for an in-depth understanding. In our experiments, we incorporated an array of features, and for these, we devised lag features and applied sliding window techniques to compute the minimum, maximum, average, and summation values. We have also added seasonality, transformation of wind direction using dummies.

Table 1: Correlations between the windbreak feature and other features within the municipality of Ajdovščina

Correlation	Feature name
0.4952	wind speed rolling min 1 day
0.4887	wind speed rolling min 12 hours
0.4412	wind speed rolling max 30 days
0.4092	mean relative humidity very high rolling sum 120 days
0.3756	wind speed 4 hours ago

5 CONCLUSION AND FUTURE WORK

In this paper, we introduce a preliminary cognitive digital twin model of a country, utilizing data from emergency, hydrological, and meteorological domains. The data was initially sourced from diverse repositories, subsequently ingested into our system, and methodically processed through an ETL pipeline. Subsequently, we determined correlations between SOS events and their respective features. Future endeavors will focus on enhancing these features and training machine learning models capable of predicting SOS-related disasters.

6 ACKNOWLEDGMENTS

The research described in this paper was supported by the Slovenian research agency, Ministry of Defence under the project NIP v2-1 DAP NCKU 4300-265/2022-9 and the European Union's Horizon 2020 program project Conductor under Grant Agreement No 101077049.

REFERENCES

[1] ARSO. 2023. Arso meteo. https://meteo.arso.gov.si/met/sl/weather/fproduct/text/. [Accessed 01-09-2023]. (2023).
[2] ARSO. 2023. Arso vode. https://www.arso.gov.si/vode/podatki/podzem_vode_amp/. [Accessed 01-09-2023]. (2023).
[3] Michael Grieves and John Vickers. 2017. Digital twin: mitigating unpredictable, undesirable emergent behavior in complex systems. *Transdisciplinary perspectives on complex systems: New findings and approaches*, 85–113.
[4] Daniel Jurgens. 2022. Creating a country-wide digital twin. https://www.wsp.com/en-nz/insights/creating-a-country-wide-digital-twin. [Accessed 01-09-2023]. (2022).
[5] Ville V Lehtola, Mila Koeva, Sander Oude Elberink, Paulo Raposo, Juho-Pekka Virtanen, Faridaddin Vahdatikhaki, and Simone Borsci. 2022. Digital twin of a city: review of technology serving city needs. *International Journal of Applied Earth Observation and Geoinformation*, 102915.
[6] Joshua C Nwokeji and Richard Matovu. 2021. A systematic literature review on big data extraction, transformation and loading (etl). In *Intelligent Computing: Proceedings of the 2021 Computing Conference, Volume 2*. Springer, 308–324.
[7] ESRI Singapore. 2023. A framework to create and integrate digital twins. https://esrisingapore.com.sg/digital-twins. [Accessed 01-09-2023]. (2023).
[8] SOS SPIN. 2023. Spin sos - uprava rs za zaščito in reševanje. https://spin3.sos112.si/javno. [Accessed 01-09-2023]. (2023).
[9] SURS. 2023. Gis. https://gis.stat.si/. [Accessed 01-09-2023]. (2023).
[10] Fei Tao, He Zhang, Ang Liu, and Andrew YC Nee. 2018. Digital twin in industry: state-of-the-art. *IEEE Transactions on industrial informatics*, 15, 4, 2405–2415.