WORLD DATA LEAGUE

Insights Report

2023

Authors: Leonid Kholkine and Tamara Fingerlin

Designer: Sara Gonçalves

The following have contributed directly or indirectly to the content of this report:

WDL Participants: Alina Carvalho, Ana Luiza Akiyama, Ana Maria Amaro de Sousa, Catarina Rocha, Clara Pedroso, David Raposo, Divya Kamat, Duarte Pereira, Duarte Rodrigues, Eduardo Vicente, Emma Roscow, Ernitia Paramasari, Frank Novak, Gonçalo Ferreira, Guilherme Caixeta, João Almeida, João Anacleto, José Almeida, José Sousa, Joydeep Chatterjee, Juliana Machado, Kate Crawford, Lara Strachan, Luiz Gustavo Serra, Madalena Diniz, Maria Castro, Mariana Amaro de Sousa, Mariana Xavier, Marta Colaço, Martim Chaves, Mine Yasemin, Mitra Ganeson, Mohamed Fouda, Pankaja Shankar, Paulo Sousa, Ricardo Brioso, Ricardo Filipe, Rodrigo Ferreira, Roisin Holmes, Rui Santos, Samantha Hakes, Sara Sabzikari, Stuart McGibbon, Xavier Silva, Zhen-Yen Chan

WDL Team: Aleksandra Borkowska, Celso Santana, Fabiana Oliveira, João Martins, Leonid Kholkine, Margarida Abranches, Miguel José Monteiro, Rui Mendes, Tamara Fingerlin

Challenge and Data Providers: Cascais Municipality, City of Ghent, LxDataLab Lisbon City Council

Financial Support: BPI Grupo CaixaBank, Daredata Engineering, Enlitia, euroclear, Fidelidade, galp, JTA - The Data Scientists, Siemens

Institutional Partner: EuroCities

Jury Members: Aruna Sri Turlapati, Bhargavi Mahesh, Catarina Belém, Fabian Tinkl-Hennighausen, Filipa Peleja, Gustavo Fonseca, Helder Oliveira, Horácio Neri, Irune Lansorena, Jacek Kustra,, Julián Miranda, Kyra Wulffert, Maedeh Afshari, Mahmoud AbdelAziz, Mariana Oliveira, Pedro Dias, Plaban Nayak, Ricardo Araújo,, Rui Braga, Shikhar Chauhan, Sofia Silvestre, Stefan Saloman, Tansif Anzar, Teresa Scholz, Tiago Otto Rodrigues

Finals Grand Jury Members: Joyca Leplae, Lambert Hogenhout, Stefaan Verhulst

Team Mentors: Afonso Oliveira, Aruna Sri Turlapati, Bhargavi Mahesh, Catarina Belém, Chanukya Patnaik, Erum Afzal, Fabian Tinkl-Hennighausen , Filipa Peleja, Floris Goes, Gonçalo Almeida, Gustavo Fonseca, Helder Oliveira, Horácio Neri, Irina Vidal Migallón, Irune Lansorena, Jacek Kustra, João Afonso Pereira, Joinal Ahmed, Julián Miranda, Kyra Wulffert, Lara Oliveira, Luís Espírito Santo, Maedeh Afshari, Mahmoud AbdelAziz, Mariana Oliveira, Nemanja Radojkovic, Paulo Maia, Pedro Chaves, Pedro Dias, Pedro Madeira, Plaban Nayak, Ricardo Araújo, Ricardo Vitorino, Rúben Menezes, Rui Braga, Rune Prytz, Shikhar Chauhan, Sofia Silvestre, Stefan Saloman, Tansif Anzar, Teresa Scholz, Tiago Otto Rodrigues, Vasco Ferreira

Executive Summary

The World Data League (WDL) strives to enable cities and communities to leverage data for social impact, adding to an ever-growing repository of open knowledge centering around the United Nations (UN) Sustainable Development Goals (SDG).

The third edition of the WDL flagship competition took place in the spring of 2023 and brought together over 100 participants from all over the world to focus on three high-impact challenges proposed by cities in Portugal and the Netherlands.

In each stage of the competition, a different SDG was in the spotlight: SDG 11 (Sustainable Cities and communities), SDG 14 (Life below water), and SDG 7 (Affordable and clean energy).

All solutions and proof-of-concept algorithms created in the competition have been made open source and are accessible on our **GitHub**, as well as in the **WDL Social Impact Hub**.

Our goal is to disseminate this knowledge openly to facilitate sustainable communities all around the globe.

In this spirit, evaluations in the WDL competition go far beyond strictly technical aspects and metrics optimization. Our evaluation matrix includes the understanding of data stakeholder's needs, concern for data quality, deep exploratory analysis, discussion of model comparison and selection, creation of a creative solution aimed at the end user and measuring the impact of the implementation, taking into account interpretability and fairness of the data product created.

In this document, the authors summarized the teams' insights and findings for each challenge, organized into five main categories:

Data describes the real-world datasets the teams worked with provided by cities. It describes what data the teams found important, where it could be improved, and what additional data would have been useful. This section aims to give an idea of what type of data might be needed to solve a certain challenge.

Methods and Techniques describes the technical aspects of the team's submissions. A short description of the types of methodologies and algorithms used is presented. This section aims to give an overview of the methodologies that could be implemented to solve a certain challenge.

Main Insights from Data sums up the interesting findings by the teams, either through data analysis or by applying certain mathematical models. This section aims to give an overview of possibilities for insights and impacts that can be achieved with little resources, as the participants only had three weeks to complete the challenges.

Product sums up how the team-developed algorithm could be used and who would use it. This section aims to showcase possible products that stem from these algorithms and identify the features, users, and outputs of those products.

Social Impact analyzes the potential impact if the organization implements the algorithm or product. This section aims to describe the desired social outcome and impact metrics.

The authors are excited to present the findings of this third successful edition of WDL and hope many cities and organizations will benefit from the data products and models created.

Glossary

Alphabetically

A3T-GCN	Attention-gated Temporal Graph Convolutional Neural network
BIMD	Belgian Index of Multiple Deprivation
EMD	Earth Mover's distance
GTFS	General Transit Feed Specification
KPIs	Key Performance Indicators
LASSO	Least Absolute Shrinkage and Selection Operator
LLM	Large Language Model
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
PCA	Principal Component Analysis
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SARIMA	Seasonal Auto-Regressive Integrated Moving Average
SARIMAX	Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors
SDG	UN Sustainable Development Goal
VAR	Vector Autoregression
WDL	World Data League

Index

Introduction

The first half of 2023 has seen numerous innovations in data science as accessible to the general public as ever. With the proliferation of Large Language Models (LLMs) packaged into chatbots, AI and machine learning discussions have found their way from data science chatrooms to family dinner tables. Data and what you can do with it are on everyone's mind, and the pace of innovation is only accelerating.

With this spotlight on the field, data scientists, even more than before, have an obligation to educate, explain, and use their skills for the social good. Data literacy and ethical use of data are now a core issue of our times, and decisions made now will shape humanity for decades to come.

However, there is a gap in expertise between organizations that have collected growing amounts of data and data scientists with the skills to help sort through the data, distill insights, and apply cutting-edge methodology to solve real-world problems.

The World Data League aims to close this gap with organizations working on socially-oriented challenges. For this purpose, we defined our challenges based on the **United Nations Sustainable Development Goals (SDGs)**.

The 2023 third edition of the WDL competition saw over **100 participants** from **17 countries** form **29 teams** to tackle **3 different challenges** over more than **4 months**.

In total, **30 technical reports** were produced, each including a code notebook, executive, and video summary of the findings and the solution that was being proposed.

How to interpret this document

In this document, we summarized the methodologies and models used, conclusions reached by the teams as well as main insights found and data products created. The authors stress that the outcomes presented here should be considered proof-of-concept with a need for scientific validation. That is due to the fact that participants were limited to the datasets presented to them (which could vary in quality, quantity, and granularity). In many cases, although there is a correlation between certain variables, it should not be considered a direct or indirect cause. The results presented here outline what the teams have presented in their reports. We hope these ideas can spark future research directions and bring new ideas on collecting and leveraging data to create social impact. All the ideas presented here can be found in the teams' full submissions on the **World Data League code repository**.

Framework for Social Impact Measurement

To achieve our mission of creating **accessible data-driven social impact solutions that can be used in real-world scenarios**, it becomes crucial to guarantee that all the solutions created by the participating teams directly consider Social Impact.

Social impact measurement cannot simply be a bonus or a nice-to-have. It **needs to be an integral part of the teams' work** and something they must bear in mind when submitting their solutions.

Therefore, we included social impact measurement as a mandatory step on which the submissions would be specifically evaluated. We believe that creating an extremely powerful solution is pointless if there is no clear path to making it usable and impactful for the social impact entities.

To make this social impact measurement easier to navigate by teams, WDL proposed a Social Impact Framework template that needed to be completed for every solution before submission. This template was distributed to the participants alongside an example.

The Social Impact Framework template includes the following items:

1. Define your **product**

- What is the input to that product? Who gives that input? Who are your "clients"?
- What is the activity of your product? What are the features? What does it actually do?
- What is the output of your product? What does it show to your "clients"? How does it show that?

2. Define your **outcome**

- If the outputs are your immediate results, what are your long-term results?
- What do you want your product to achieve?
- What "good" are you creating?

3. Define your **impact metrics**

- From the outcome, define **3 to 5 metrics that actually measure** if you are achieving that outcome or not.

4. Estimate **how much those metrics** will change with your product deployment

- Since you cannot wait to see the impact of your product, **estimate it** using the estimations/predictions of your model, market research, products from proxy industries or location

WDL 2023 Topics

In 2023 we are living in a world facing many challenges related to the wellbeing of humans and their ecosystem, which often are interrelated and systemic. At the same time, smart technologies, sensors, and open data initiatives, in combination with advances in the field of data science, allow new data-driven solutions to these vital challenges.

In the first edition in 2021, WDL focussed on Sustainable Cities and Communities (SDG 11). After the success of opening the competition to embrace challenges targeting other SDGs in 2022, we decided to allow cities to enter challenges aiming at the social good flying under the flag of any SDG.

Contrary to the previous year and to maximize the focus, participants were given only one challenge per stage.

The three challenges that were selected centered around:

Phase 1:

Mobility
(SDG 11 - Sustainable Cities and Communities)

Semi-Finals:

Biodiversity
(SDG 14 - Life below water)

Finals:

Energy
(SDG 7 - Affordable and clean energy)

Phase 1

Mobility

Determining The Main Mobility Flows in the City of Lisbon Based on Mobile Device Data

Challenge by
LxDataLab at the Lisbon City Council

LISBOA
CÂMARA MUNICIPAL

In the last decades, the city of Lisbon has observed a loss of inhabitants from the downtown to its metropolitan area. From the early 1980s until 2017, the number of inhabitants of the city decreased from 800.000 to 500.000. Simultaneously, car use in daily commuting between the city and the metropolitan area showed a clear increase.

This has resulted in an overload of the road network and parking spaces in the city and a decrease in safety and quality of life for the city's inhabitants and users. In 2017 use of public transport decreased from about 46% of journeys that start and end in the city, compared to only 22%.

Reversing the current modal split to free up public space for citizens and ensuring convergence with the goals of the Paris Agreement, namely carbon neutrality by 2050, is the biggest challenge of the mobility policies for the city of Lisbon. It is thus necessary to change the paradigm.

The city of Lisbon proposed this challenge to get a better understanding and visualization of how people move between grids during rush hours (7:00 AM - 10:00 AM and 5:00 PM - 8:00 PM) and a model that can predict those movements and identification of potential interventions to improve the commuting experience of people in Lisbon and favor sustainable modes of mobility.

For this challenge, the LxDataLab team stressed the importance of a predictive model as the core desired outcome and welcomed more general and niche solutions to aspects of the problem posed.

Goal

This challenge aims to extract the inputs that will allow the planning and execution of the necessary actions to improve mobility in the city of Lisbon, the quality of life of its citizens and meet sustainability goals.

United Nations SDG

- GOAL 11: Sustainable Cities and Communities
  - Target 11.2.1: Provide access to safe, affordable, accessible, and sustainable transport systems for all.

Datasets

- Number of mobile phones entering, remaining, and exiting per 200m/200m square in a period of 15 minutes - Lisbon City Grid (September to November 2022)
- Number of mobile phones entering and exiting the city every 15 minutes on the 11 main axes of entry into the city of Lisbon - Axes of the city of Lisbon (September to November 2022)
- Identification of the 11 points of entry and exit of Lisbon
- Data on the road network of the city of Lisbon
- Traffic level data (from the WAZE platform) and traffic conditions

Data

This challenge came with a collection of large, highly granular, and well-documented datasets already offering many possibilities for different approaches and the development of different data products. Most teams worked on a selection of the datasets provided and integrated their selection with relevant external publicly available data to get more insights into specific issues.

Several teams supplemented the datasets with data from Lisbon's Open Data portal **Conjuntos de dados** on public transport, road networks, environmental and meteorological information. **One team** added bus data **(GTFS)** as a proxy for public transportation, another team used data on the cycling infrastructure from **OpenStreetMap** and yet another team decided to collect points of interest in Lisbon such as restaurants from google maps.

Methods and Techniques

One team decided to focus on traffic jams and benchmarked several time-series models on custom metrics they had derived from the provided datasets. This team defined a *Flow* metric measuring the movement of individuals within grid cells in Lisbon and a *Jam level* metric that assessed traffic intensity within grid cells.

The models this team evaluated to predict traffic jams included Last Value Repeating & Last Cycle Repeating naive benchmarks, a neural network based on dense layers as well as a recurrent neural network (RNN) trained on the whole history of the training data using Long Short-Term Memory (LSTM) layers.

Understanding peak rush hour movement patterns was the focus of **another team**. They used the **OpenCV library** to analyze movement between grids with the Wasserstein distance / EMD (earth mover's distance), inspired by **Balzotti et al. (2018)**. Their predictive model of choice was **NeuralProphet**, a hybrid forecasting framework based on **PyTorch**. The very same team also employed a **LISA / Local Moran map** to detect spatially extended clusters and diagnose local instability (Figure 3) in the data on distinct terminals in grid squares during rush hour.

A **third team** focussed on predicting key flows of people currently occurring in the city of Lisbon using an attention-gated temporal graph convolutional neural network (**A3T-GCN**). They explained their choice to combine the analytical capabilities of graph CNNs for spatial relationships with gated recurrent units, a commonly used layer in complex temporal modeling.

Main Insights from the Data

One team focussed on predicting the overall number of traffic jams over time, which they visualized using **plotly** (Figure 1).

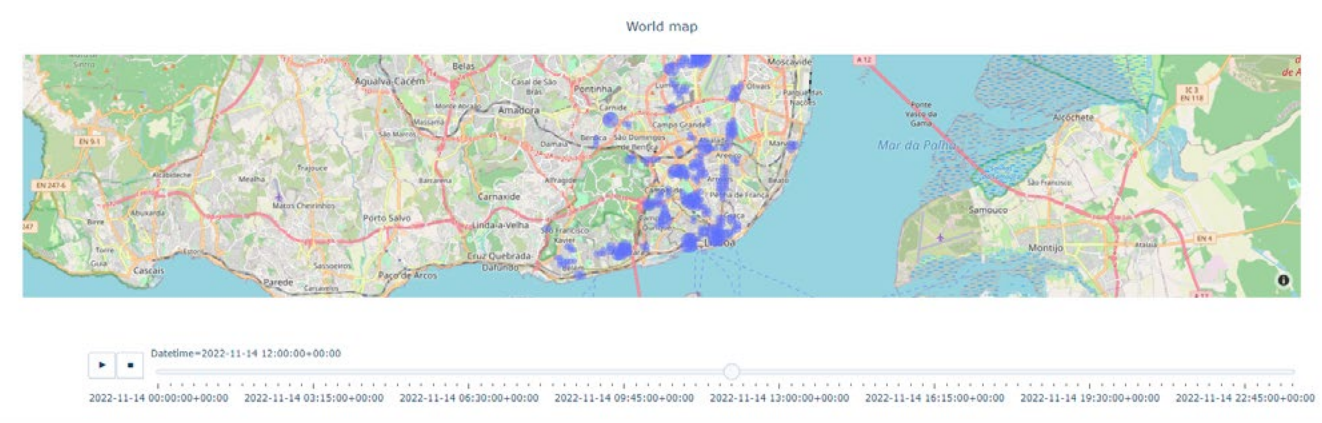


Figure 1. Visualization of the number of traffic jams in Lisbon over time

They noted the cyclical nature of the traffic impact of rush hours shown by the high performance of the Last Cycle repeated benchmark and found the most successful approach given the available data to focus on time-based features only and deploying an LSTM model (Figure 4). This model reached a **performance** of 0.24 loss MSE / 0.31 metric MAE, clearly outperforming the baseline models.

The team focussing on rush-hour movement patterns performed extensive exploratory data analysis and visualization using **hvplot** and **geopandas** in addition to plotly (Figure 2).

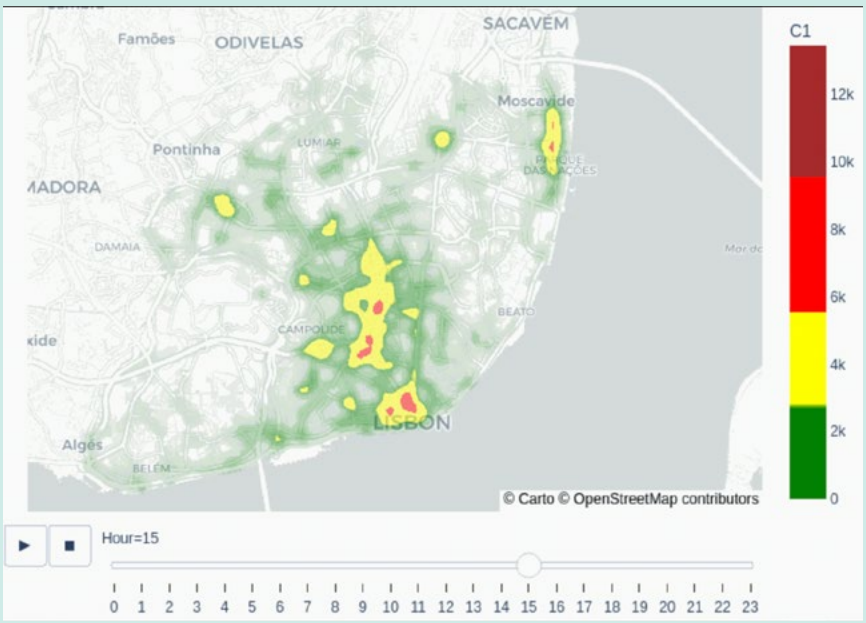


Figure 2. Heatmap of the average number of distinct terminals in the grid over time

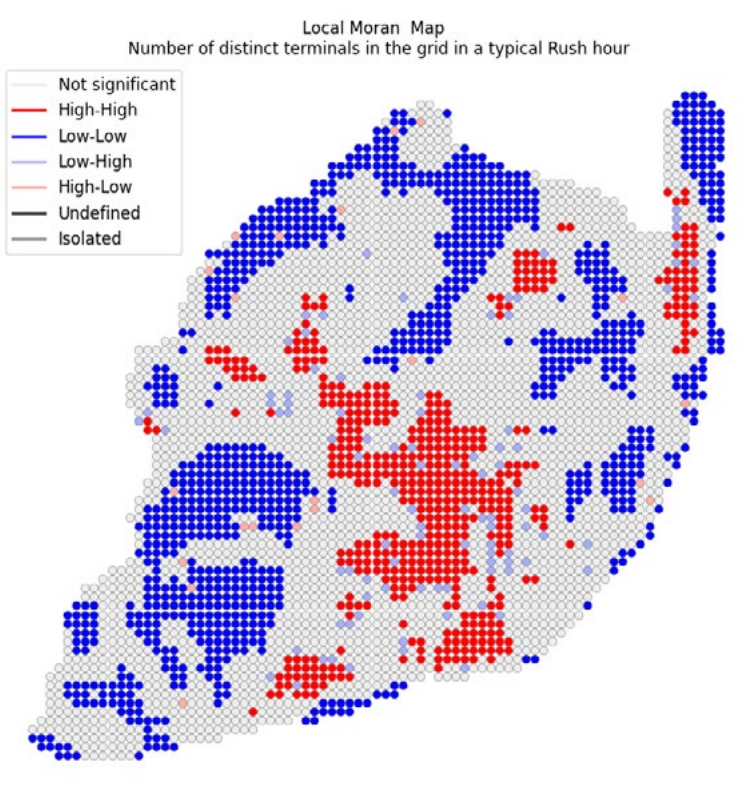


Figure 3. A Local Moran map of Lisbon using the LISA cluster analysis on the density of distinct mobile phone terminals in a grid during a typical rush hour. Red areas show "high-high" groupings, meaning these grids showed high values and were surrounded by high values; conversely, blue areas are "low-low" groupings, areas with few mobile phones surrounded by areas with few mobile phones.

In their EDA, this team determined that the Avenidas Novas, Arroios, São Domingos de Benfica, Santo António, and Parque das Nações neighborhoods show the highest people density during rush hours and the highways with the most incoming and outgoing traffic were IC19, IC2 and IC16, traffic jam occurrence generally correlated with areas of high terminal density. In general, the flow of people appeared directed towards and from the very center area of Lisbon and seemed to mainly stem from Portuguese phone numbers during common commuting times.

Product

The products proposed by teams were as diverse as their approaches.

One team designed a tool for urban planning professionals aiming to predict high traffic impact (Figure 4) and possible effects of changes in transportation infrastructure and approaches. Users with no modeling experience would be able to quickly assess and visualize the impact of proposed traffic interventions guiding decision-making and governance.

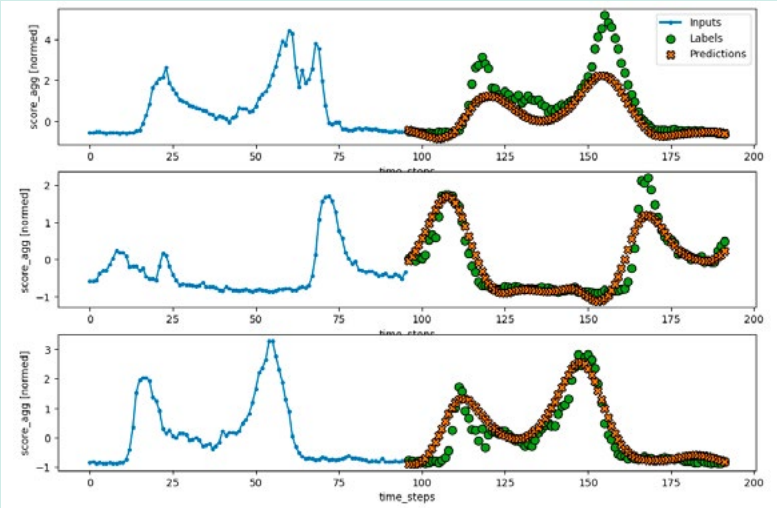


Figure 4. Testing predictions for traffic jams in Lisbon with an LSTM model 4.4.3.1.

An analytics web app, "Peak Analytics", was proposed by **another team** with the goal to help forecast the movement of people during events that commonly cause increases in traffic like music festivals or soccer games. This product was aimed at government officials to compare plans with past events and get predictions on expected traffic and air in the planning of reallocating buses and when to incentivize commuters to travel off-peak in public transport.

Another team focussing on analyzing existing traffic flows proposed to create a dashboard available to the public to assist in individual commute planning, giving agency to the citizens of Lisbon and providing the tools to avoid high flow times and areas. They also floated the idea of gamification, incorporating their insights into a traffic app and giving points to users who choose to travel less crowded roads.

Social Impact

Following the city of Lisbon's overall goal of promoting more sustainable modes of transportation and improving the quality of life of its citizens, the common denominator of the products proposed was a shift from individual to public transportation as well as a reduction of traffic-related friction like traffic jams in the city.

One team elaborated that predicting traffic jams could aid in planning the intensification of public transit routes. By providing targeted public transit, **CO2 emissions could be reduced by 20% per passenger km**. Additionally, a predictive model could help identify the ideal locations to implement **traffic flow measures**, reducing the time spent in jams, reducing CO2 emissions and improving quality of life and air.

Being able to predict traffic flows due to events using a tool like "Peak Analytics" could both lead to improved extra availability of public transportation and to understanding traffic load during rush hours in general, aiding in creating campaigns to shift commuters to travel in off-peak periods using incentives like random rewards, personalized offers.

A publicly available dashboard or app could help Lisbon residents and tourists to make informed decisions that make their commutes efficient and reduce travel time spent per individual user.

Semi-Finals

Biodiversity

Avencas Marine Protected Area: Predict the future of the local ecosystem and its species

Challenge by
Cascais Municipality

CASCAIS

The Avencas Marine Protected Area (AMPA) is a Biophysical Interest Zone in Cascais, Portugal. The AMPA has been under close observation since 2010, with regular biodiversity sampling taking place and the source of a case study by **Ferreira et al. (2017)**.

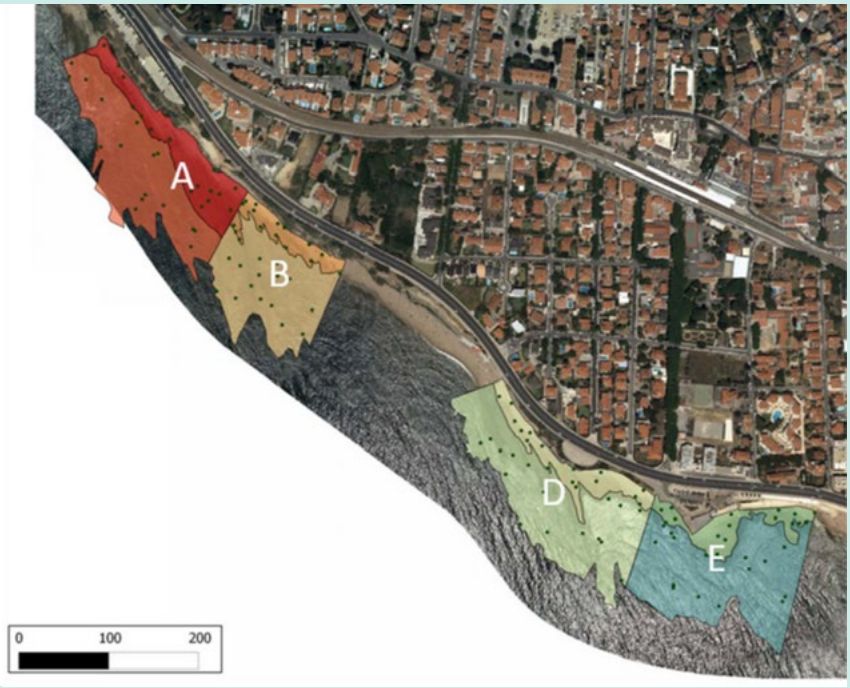


Figure 5. Location of the sampling areas in the AMPA.

To help protect this unique marine ecosystem, measures were taken to reduce human interference, **but the system did not recover as well as expected**. This is why the municipality of Cascais is looking for help getting a long-term analysis of changes in the abundance of species in the AMPA.

The two main focus areas are:

1. Examine potential factors that influence the abundance of species in the AMPA and could cause the lack of recovery of biodiversity. The city of Cascais is looking for out-of-the-box ideas of potential correlating factors.

2. Determine the trajectory species (especially endangered and invasive ones) are on in the AMPA to help provide evidence for the usefulness of the current measures if the development is positive or for advocating for more protection if the development is negative.

Goal
The goal of this challenge is to identify variables that potentially impact the marine ecosystem of the Avencas Marine Protected Area and predict further developments with a special focus on endangered and invasive species.

United Nations SDG

- GOAL 14: Life below water
  - Target 14.2 By 2020, sustainably manage and protect marine and coastal ecosystems to avoid significant adverse impacts, including by strengthening their resilience and taking action for their restoration in order to achieve healthy and productive oceans.
  - Target 14.3 Minimize and address the impacts of ocean acidification, including through enhanced scientific cooperation at all levels

Datasets

- Percentage coverage with sessile species in samples taken between 2011 - 2020 in the AMPA.
- Number of mobile species in samples taken between 2011 - 2020 in the AMPA.
- A reference list of which species are considered invasive and which are considered endangered according to the IUCN. (Note that for some species, the Portugal-specific conservation status assigned by the IUCN is given)
- Bathymetric data of the AMPA.
- Sampling area shapefiles.

Data

Cascais' challenge centered on identifying new variables that might be impacting the marine ecosystem of the AMPA. Due to the exploratory nature of the challenge, teams used many additional data sources. The following list is a selection.

- **Land, sea and air monthly anomaly temperatures: University East Anglia Climate**
- **Data**
- **Historical weather from Open Meteo**
- **Historical weather from Meteo Blue**
- **Marine Geochemical Data**
- **Air Quality Index Data**
- **Ocean Health Index Data**
- **OceanSODA-ETHZ**
- **Daily Fishing Effort at 10th Degree Resolution by MMSI, version 2.0, 2012-2020**
- **Fishery statistics from the Instituto Nacional de Estatistica**
- **Endangered Species red list**
- **Index of Coastal Eutrophication**
- **Ecological Marine Unit Explorer**
- **Biodiversity in Portugal**
- **EU reporting of Industrial Water Pollution**

One team suggested setting up a measurement station close to the AMPA to gather data on chemical and physical properties directly in the area of interest.

Methods and Techniques

Approaches of the teams for this channel ranged from using SARIMA, LASSO, gradient boosting regressor, XGBoost and random forest regressor models to determine features importance. For predictive modeling, several teams relied on time-series models like SARIMA and VAR.

For data preprocessing **one team** leveraged PCA to reduce the dimensions of the environment data.

One team developed SARIMA models to predict the number of invasive species, the number of endangered species and the **Shannon-Wiener Index** derived from the mobile abundance data and a converted version of the sessile data using the methodology of **Deepananda and Macusi (2013)** with the formula:

H=-Σ[(pi)×log(pi)]
H=Shannon diversity index
pi=proportion of individuals of ith species in the population

Fitting a SARIMA model to the derived Shannon-Index allowed for residual analysis to determine feature importance. This team chose to use a SARIMAX model incorporating the features identified for use in their product with an RMSE of 0.20 and an MAE of 0.17.

Another team determined diversity by using the Hill-Simpson metric inspired by **Roswell et al. (2021)**. After testing several models, the team selected a LASSO model based on its strong regularization to determine feature importance. They noted that none of the many weather-based variables they analyzed was a strong predictor of species diversity but there was a trend that lower water temperatures, humidity, precipitation, water vapor pressure deficit, and a higher cloud cover was associated with more biodiversity.

Main Insights from the Data

One team noticed how well the biodiversity in AMPA correlated with Ocean health indexes for all of Portugal postulating that this might indicate that the reason for the slower recovery of species in the AMPA could be caused by global rather than local factors. Invasive species were correlated with higher chlorine levels and endangered species by temperature-based features.

In general, the biodiversity seemed stable over the years, which correlated with the information from the domain experts, that they did not see the recovery expected by their interventions.

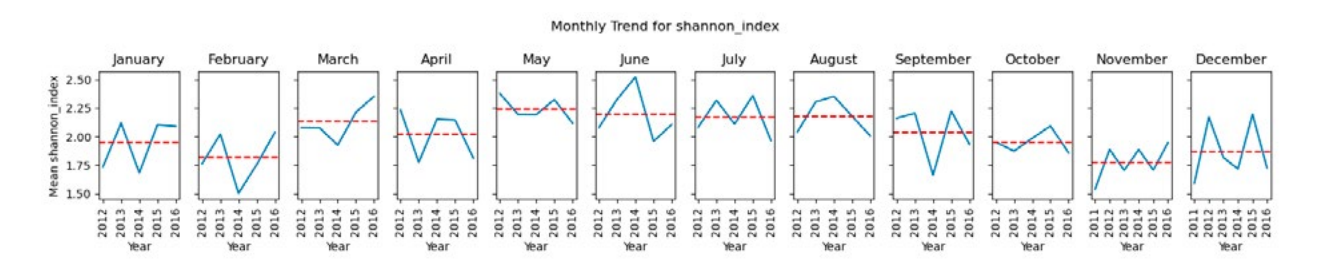


Figure 6. Analysis of one team of the monthly trend for the shannon index, showing a stationary biodiversity with strong inter-month differences.

Several teams noted that the occurrence of the vast majority of species was rare with only a few species being common in many samples, posing a difficulty for modeling.

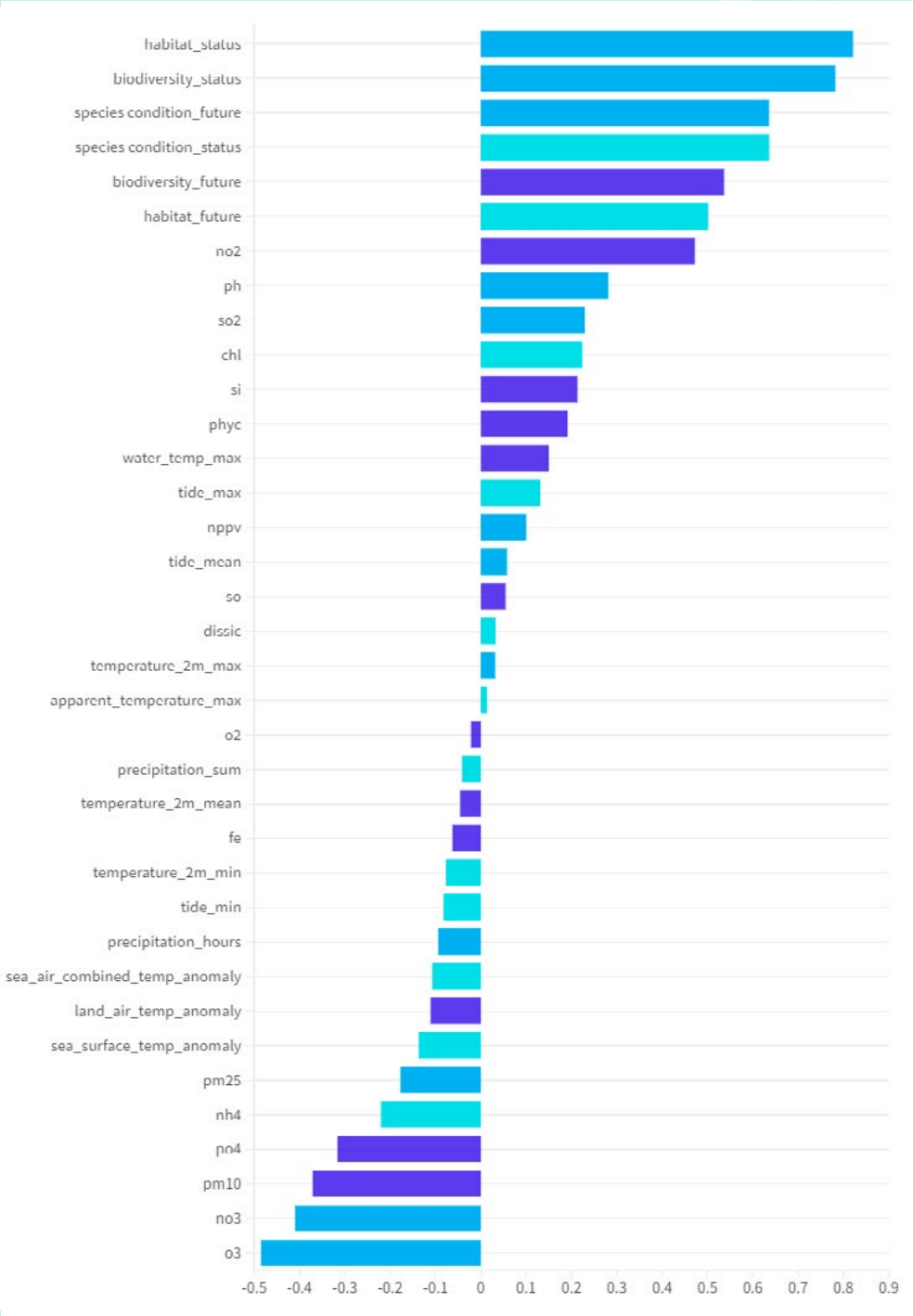


Figure 7. Feature selection based on time series model residuals by The Bayes Bunch

After testing conventional models, **another team** decided to add a more complex approach using an LSTM followed by fully connected final layers both with and without the 5 features: tide, weather condition, water temperature, season and moon phase trying to predict the abundance of Cladophora sp. Smooth, a green algae (Figure 10). They noted that adding these 5 features while increasing the train and validation performance did not meaningfully increase the performance of the model on the test set.

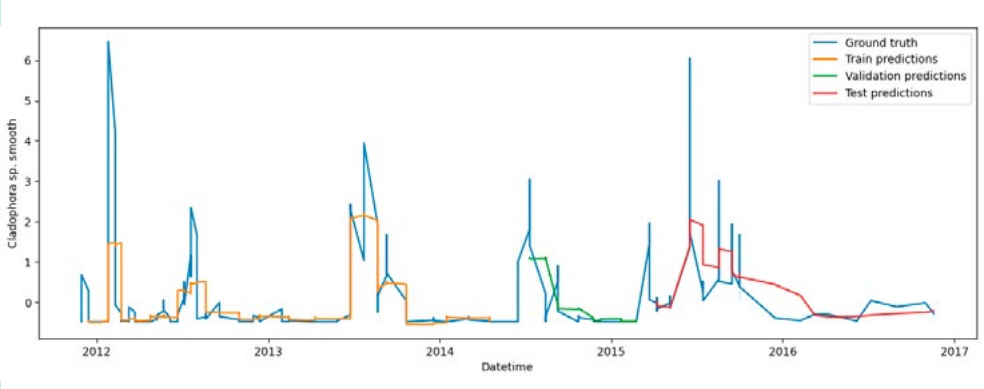


Figure 8. Predictions of an LSTM model trained with 5 extra features: tide, weather condition, water temperature, season and moon phase on the abundance of Cladophora sp.

Product

One team incorporated both their feature selection and forecasting work relative to ocean pH in a **dashboard built with Streamlit** (Figure 9).

Additionally, this team created an open-source Python package **beautiful-sea** aimed at scaling their findings to other marine ecosystems.

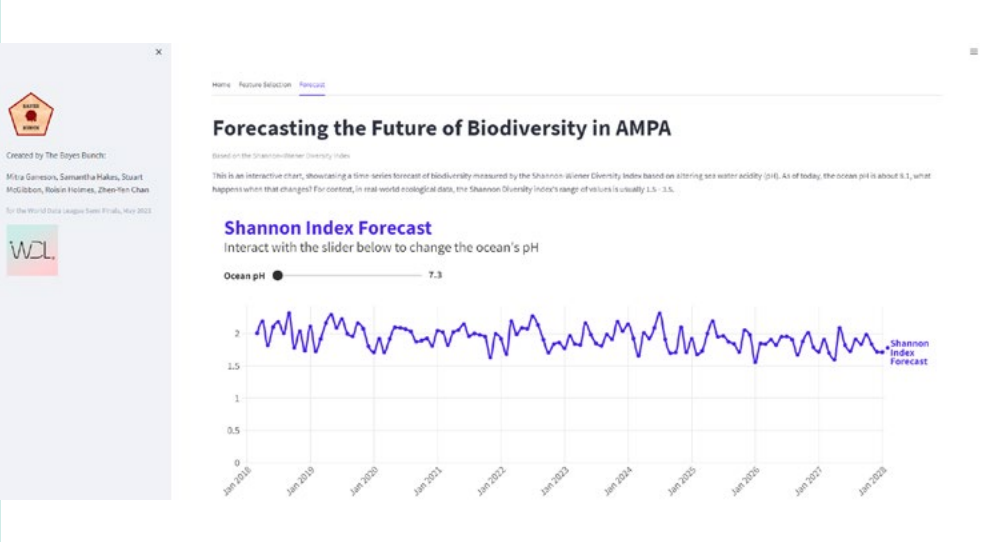


Figure 9. Streamlit App showing forecasting of the Shannon Index based on ocean pH.

Another team created a dashboard showing many of their findings such as feature importances derived from a catboost decision tree algorithm and abundance data for individual species as well as their conservation status.

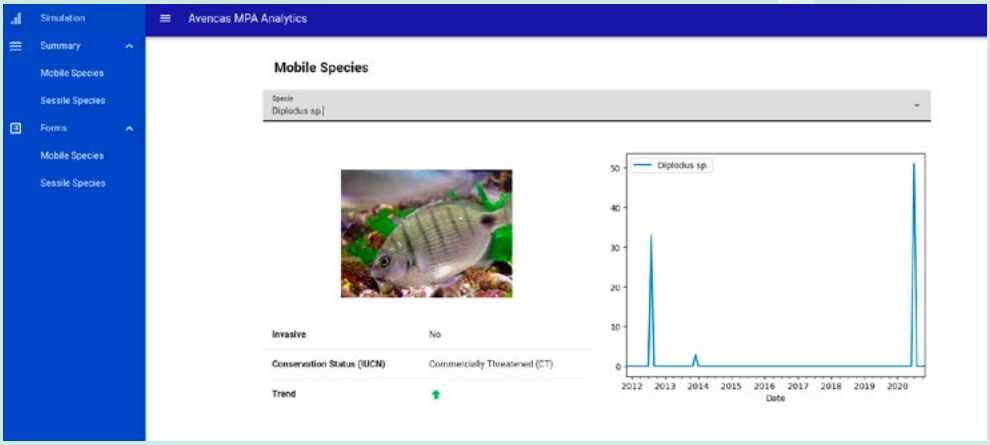


Figure 10. Predictions of an LSTM model trained with 5 extra features: tide, weather condition, water temperature, season and moon phase on the abundance of Cladophora sp.

A **third team**, after noticing that invasive species seem to thrive more when the ocean is getting warmer, looked into existing technology which could lower the sea temperature and proposed using shade balls. While shade balls are controversial in their initial purpose to save water due to requiring **a lot of water to be manufactured**, this team proposed this alternative use case for them in the interest of biodiversity. In any case, the focus on sea level temperature is extremely relevant since as of writing this metric has been off the charts with yet unknown impact on biodiversity and marine as well as all ecosystems.

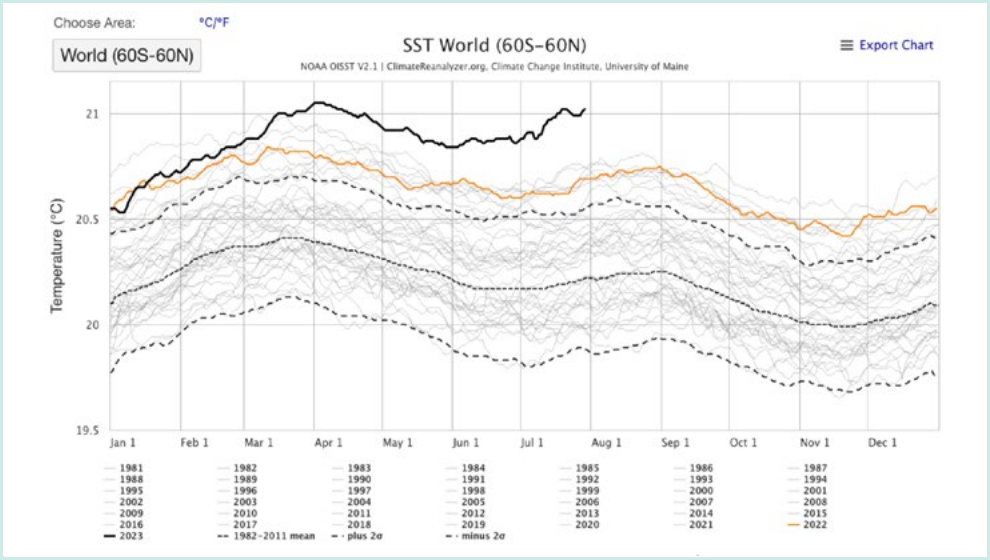


Figure 11. Global Sea surface level temperatures showing unprecedented values in 2023.

Social Impact

Using the insights from the teams, researchers and other interested parties are able to examine features for which modeling showed a high correlation with biodiversity measurements for their biological plausibility. In a second step this work can be used as a basis for specific interventions aimed at protecting the AMPA and to raise public awareness for marine biodiversity, especially in the local population and tourists visiting the area.

One core finding was the correlation between ocean acidification and a lower Shannon diversity index, indicating less biodiversity. This is especially prudent because climate change is **causing the ocean to acidify**, which is a potential direct link between the climate crisis and biodiversity in a local and well-studied marine area. Highlighting this connection to tourists and locals could increase climate change awareness and action by making the local impact tangible.

Finals

Energy

Energy communities inclusive of residents vulnerable to energy poverty

Challenge by
City of Ghent

Ghent is the second largest city in Flanders, Belgium, with a population of approximately 260'000. The City of Ghent has huge ambitions in the fields of climate mitigation and adaptation, including making clean and renewable energy accessible to all its citizens.

Energy communities are an effective way to allocate and distribute energy equitably. Recently the EU launched the **Energy Communities Repository**, defining different types of energy communities with the core tenant of their primary purpose being to generate social and environmental benefits rather than financial profits.

The city of Ghent has collected an extraordinary dataset of its solar potential using laser measurements during several flights in 2013, creating a point cloud from which their data team calculated a solar potential map spanning the whole city.

In combination with data on energy consumption, existing local energy production, and demographic data, Ghent is looking to use this dataset to create energy communities in Ghent with a special focus on including residents vulnerable to energy poverty.

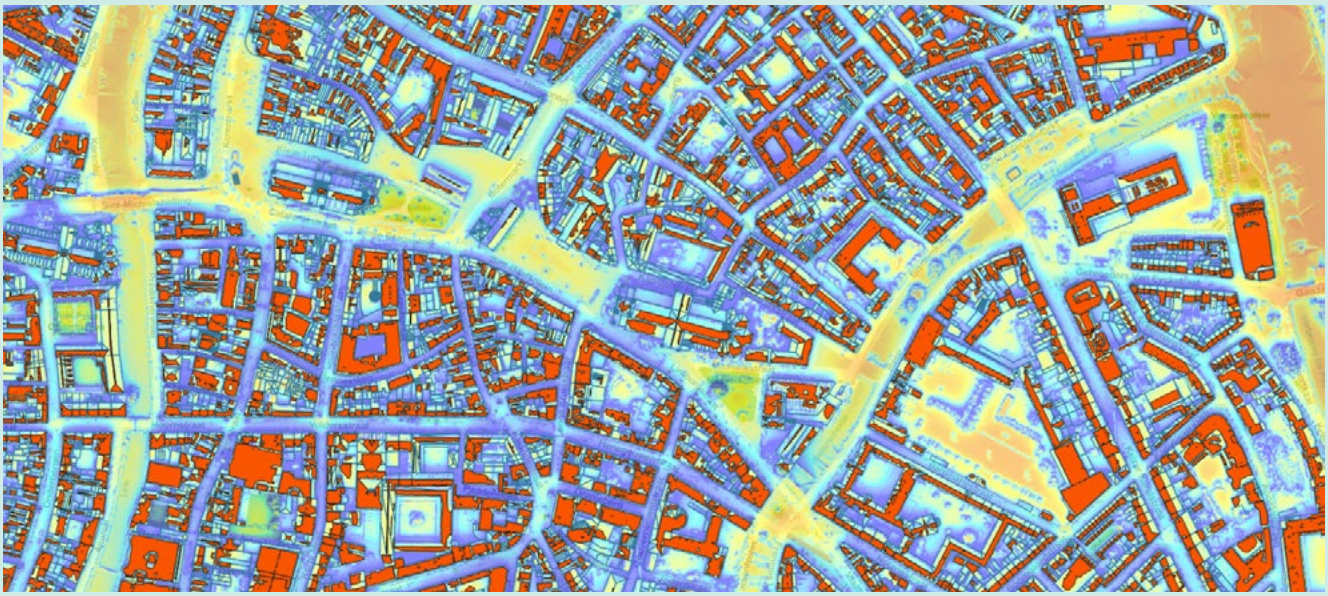


Figure 12. A small section of the solar potential map of Ghent is provided on their website.

Goal

The goal of this challenge is to leverage the existing data to propose how energy communities could be formed in the city of Ghent with a special focus on including residents vulnerable to energy poverty.

United Nations SDG

- GOAL 7: Affordable and clean energy
  - Target 7.1: Ensure universal access to affordable, reliable and modern energy services
  - Target 7.2: Increase substantially the share of renewable energy in the global energy mix

Datasets

- Geographical data on sun irradiation on a 3D model of Ghent which can be explored on the **city of Ghent's website**.
- The energy company data on energy consumption in Ghent.
- The energy company data on production, storage devices, and EV charging points.
- Ghent average income per sector.
- Current rent prices in different locations in Ghent.

Data

This year's final challenge came with a comprehensive dataset about the sun irradiation and solar energy potential on all roofs in the city of Ghent, derived from a 3D model based on a point cloud gathered from 16 points per m2 point cloud.

One team decided to augment that dataset with data on hours of sunshine in the city of Ghent, as well as installation costs of solar panels and energy tariffs in Belgium.

Another team discovered the recently published Belgian Index of Multiple Deprivation (BIMD) by **Otavova et al (2023)** and incorporated it into their model alongside open street data and demographic information provided by the city of Ghent.

Methods and Techniques

The city of Ghent posed their challenge as an open problem, looking to find ideas and ways. In order to define their data product **one team** defined their own key performance indicators (KPI), such as annualized profitability for each household in the energy community, start-up capital needed and geographical proximity of members and combined them into one outcome measurement.

Several teams (1, **2**) achieved clustering using K-means, which is well suited to a problem wanting to minimize the distance between cluster members. **One team** derived a custom iterative tiered optimization algorithm the team developed. The models were compared to a random assignment baseline.

Another team in additional to K-means tried

- spatial optimization algorithms with **queen contiguous weights** and **max-p regionalization**,
- Spectral clustering
- HDBSCAN
- as well as reinforcement learning

arriving at a combined method with rule application before clustering using K-means for their MVP (minimal viable product) in order to optimize for less compute time needed.

Main Insights from the Data

One team showed that their predicted energy communities would on average save each member of an energy community 1000 euro per year. They also pointed out that higher returns often are connected to higher upfront costs, meaning that to serve households at risk of energy poverty a tradeoff has to be considered.

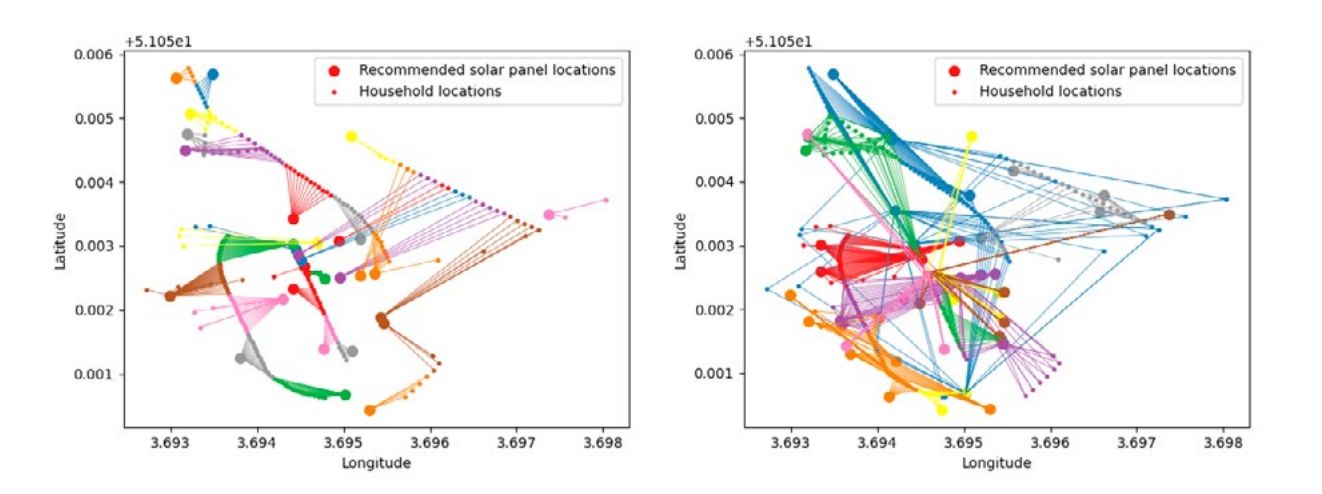


Figure 13. Plots of clustering results on a subset of the data. The left plot shows the results of K-means clustering scoring at 2621.7 in the team combined KPI metric. The plot on the right side shows the proposed energy communities cluster created by their iterative algorithm, scoring 3398.5 in their combined KPI metric, both models meaningfully beat the random assignment baseline.

Another team plotted the area with the highest solar potential, showing it to be highest around the city center, where individuals with higher risk to energy poverty were grouped as well (Figure 14).

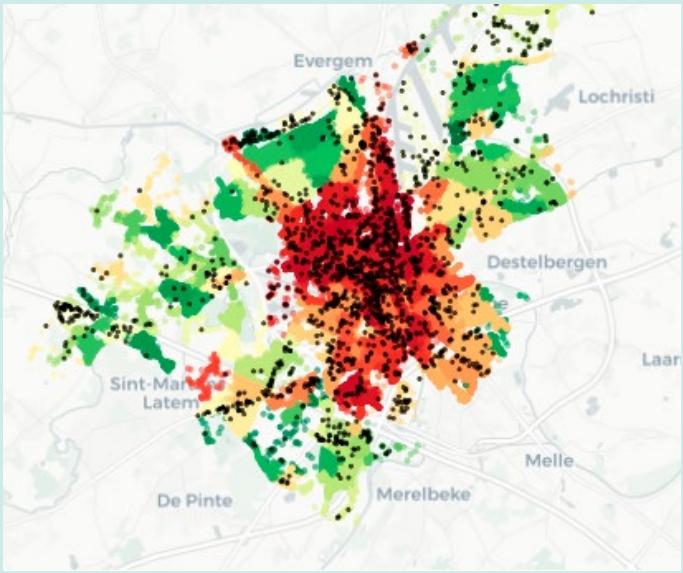


Figure 14. Roofs with the highest solar potential as black dots, ranked income domains as colored areas (red being the lowest income on average).

Product

One product proposed was a web app open to the public that can give recommendations to the inhabitants of Ghent with whom and how to start an energy community in their neighborhood. The app highlights both potential financial gain, as well as a reduction in carbon emission per household (Figure 15).

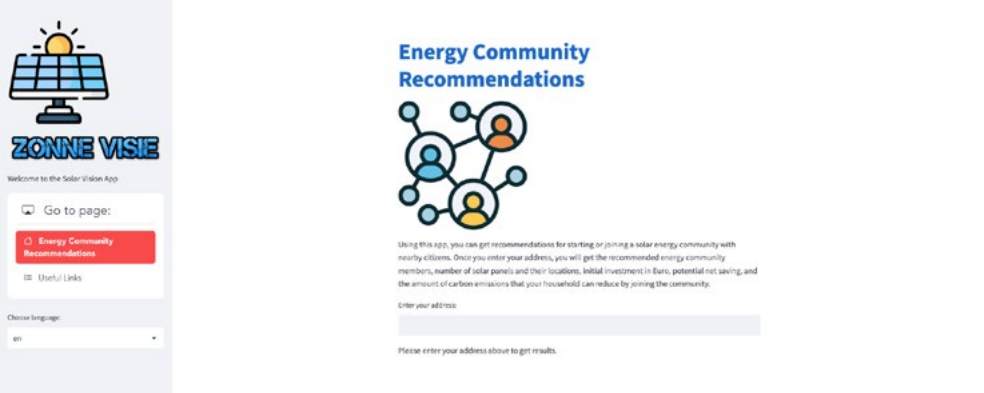
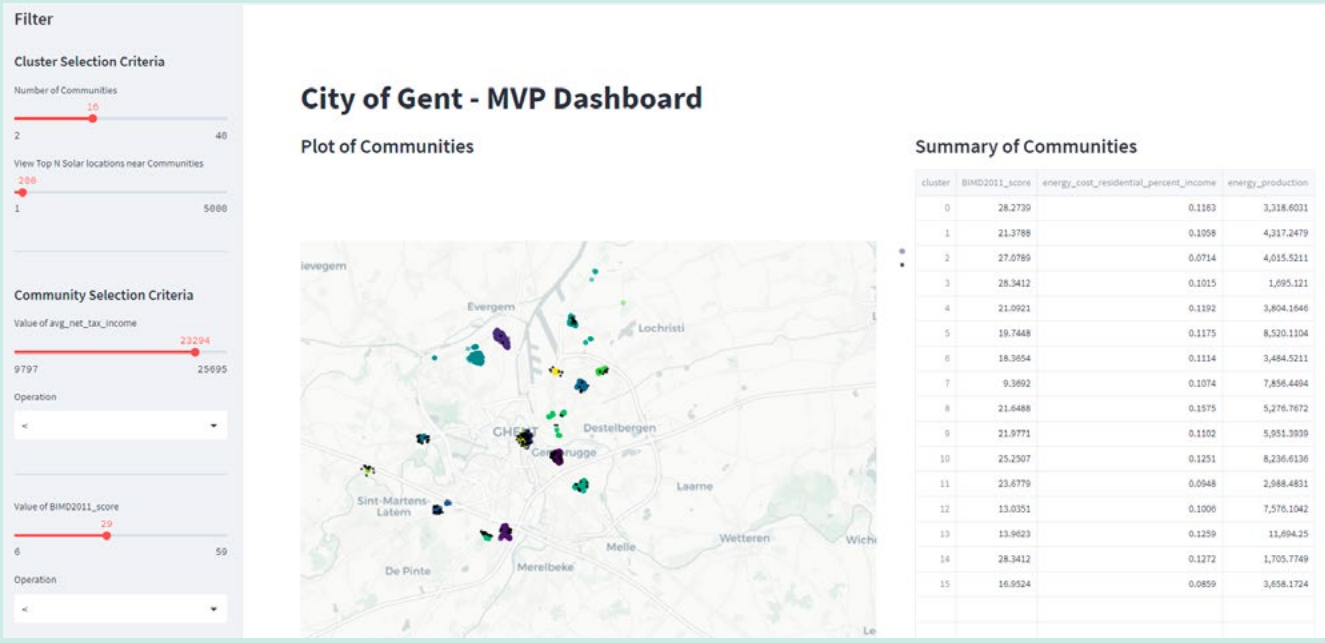


Figure 15. Web app of one team showing the benefits of an energy community to the public.

Another team developed their own **open source Python package** to be used to optimize the formation of energy communities. In addition they created a dashboard for non-technical users to explore energy communities suggested based on user-defined inputs.



cluster	BIMD2011_score	energy_cost_residential_percent_income	energy_production
0	28.2739	0.1163	3,318.6031
1	21.3788	0.1058	4,317.2479
2	27.0789	0.0714	4,015.5211
3	28.3412	0.1015	1,695.121
4	21.0921	0.1192	3,804.1646
5	19.7448	0.1175	8,520.1104
6	18.3654	0.1114	3,484.5211
7	9.3892	0.1074	7,856.4494
8	21.6488	0.1575	5,276.7672
9	21.9771	0.1102	5,951.3939
10	25.2507	0.1251	8,236.6136
11	23.6779	0.0948	2,988.4831
12	13.0351	0.1006	7,576.1042
13	13.9623	0.1259	11,694.25
14	28.3412	0.1272	1,705.7749
15	16.9524	0.0859	3,658.1724

Figure 16. A second team approach to a dashboard showing different energy communities based on user-defined values.

A **third team** applied a K-means algorithm to the whole dataset solving for the problem of optimal energy communities if all individuals living in Ghent should be assigned to one (Figure 17).

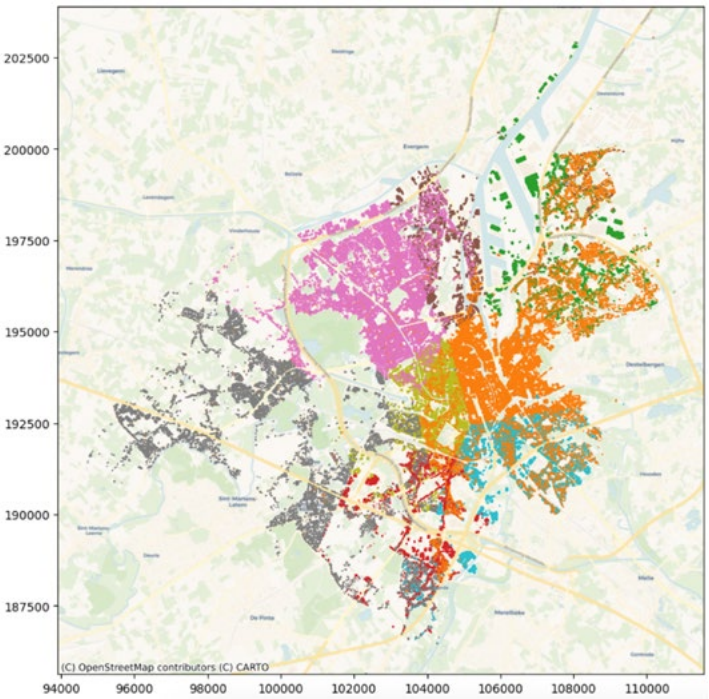


Figure 17. Plot of K-means clustering including all households with the ideal K of 8 determined with the Elbow method.

Social Impact

When analysing the potential impact of their solutions several teams noted that the challenge proposed is as much a technical as a social one. Helping residents see and understand the environmental and financial benefits of being part of an energy community is key and secondary benefits like social cohesion and empowerment of citizens

With data gathering already in place concrete impact metrics can be reduction in energy consumed as well as reduction of carbon emissions. Previous analysis by **Felice et al (2022)** on the potential for energy communities in Belgium estimated a potential cost-reduction of 10-26% and emission reductions of 5-13%, with the most important challenge being to reduce the barrier to entry for citizens, which was one of the core aspects addressed by proposed solutions.

Conclusions

In this report we summarize countless hours of work of over 100 participants on three technically and creatively challenging data problems. At the end of this over three-month-long competition we are proud to have 30 unique solutions to three challenges, each solution complete with an executive summary, a full notebook, and a video pitch. All results of the competition are open-source and can be found in the **wdl-solutions GitHub repository**, alongside the leaderboard showing which solutions were rated highest by our expert jury.

One of the main goals of WDL is to produce novel solutions to real-world problems using real-world data by cities to allow for data-driven improvement in these very cities. This creates an additional challenge for participants, working with real data in all its complexity and imperfection. We think the participants of World Data League 2023 rose to this challenge and showed us and the cities involved applicable and holistic proof of concept solutions we hope will be developed further in the months to come.

Not only that but by keeping scalability in mind we aspire to continue to add to our repository of solutions, the **Social Impact Hub**, to aid not only the cities that provided the solutions but any entity that has similar data and is looking to leverage it to advance SDGs in their area.

Follow us on **LinkedIn** and **YouTube**.

Interested in our work or solutions?

hello@worlddataleague.com