WORLD DATA LEAGUE

Insights Report

2022

Authors: Miguel José Monteiro and Leonid Kholkine
Editor: Fabiana Oliveira
Designer: Celso Santana

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

WDL Participants: Abdelrhman Elkhouly, Ahmad Elkholi, Ajit Gupta, Akshay Punjabi, Alessandro Consiglio, Alexander Schiel, Alexandra Alexandru, Alexandra Serras, Alfredo Petrella, Amit Sahoo, Ana Fouto, Ana Luiza Akiyama, Anant Pingle, André Luís, Aníbal Silva, Antonio Oliveira, Any Pereira, Artur Amorim, Beatriz Barreto, Beatriz Lourenço, Benjamin Chew, Bianca Dragomir, Bruno Alho, Bruno Ribeiro da Silva, Catarina Bento, Cátia Correia, Chiara Rucco, Claire Benard, Claudia Cozzolino, Cristian Castro, Cristiana Carpinteiro, Dagoberto José Herrera Murillo, Daniel Tan, David Melo, David Gamba, Diego Arenas, Diogo Baptista, Diogo Pessoa, Diogo Polónia, Elena Viganò, Emil Henriksson Ene, Emil Pedersen, Emily Martins, Enrico Coluccia, Fábio Lopes, Fernando Chavez Osuna, Filip Claesson, Francisca Ruiz-Pérez, Francisco Valente, Frederico Rodrigues, Freth Arguedas, Gabrijela Juresic, Gonçalo Carvalho, Gonçalo Costa, Gonçalo Salgueirinho, Guilherme Caixeta, Guisela Arguedas, Hayanne Oliveira, Heloise Rozier, Hiba Laziri, Hongyu Ao, Houman Heidarabadi, Imre Boda, Irune Lansorena Sanchez, Isabella Olariu, Israel Souza, Ivan Vrkic, Joana Camargo, João Afonso Pereira, João Almeida, João Matos, José Castro, José Cerqueira, José Ferreira, José Mauricio Nunes de Oliveira Jr, José Sá, Juan Diego Arango, Júlia Schubert Peixoto, Julian Jensen, Júlio Medeiros, Karim Anwar, Kenneth Goh, Kirti Tyagi, Kylix Afonso, Lúcia Moreira, Luckshan Sivakumar, Luis Ventura, Luiz Gustavo Moniz Serra, Luiza Corpaci, Lukas Böhm, Lukas Mölschl, Mads Hoefer, Magdalena Brach, Mai Chi Do, Manuel Borges, Manuel Bürgler, Marc Behse, Marcelo Moreno, María Rodríguez-González, Mariana Martins, Marta Seca, Martim Chaves, Melissa Montes Martin, Miguel Zina, Mohamed Al Sayed, Mohamed Elshirif, Mohamed Nabil, Mohamed Taha, Mónica López-Lacort, Natalie Muenter, Natascia Caria, Navid Safari, Neha Shaah, Nicholas Sistovaris, Nikhil Kulkarni, Nuno Lavado, Pablo Izquierdo Ayala, Patrícia Rocha, Pedro Costa, Pedro Fernandes, Pedro Fernandes, Pedro Fonseca, Pedro Leal, Pedro Miguez, Pedro Ruas, Peter Michaletzky, Renata Costa, Robert Nyquist, Romane Le Goff, Roney Mathew, Rui Granja, Rui Monteiro, Sai Pravallika Myneni, Sandra Martínez-Sanchis, Santiago Cardona Urrea, Shiv Yucel, Sofía Ramírez, Sourabh Hujare, Tejas Choudekar, Terry Zhang, Tiago Neves, Tin Wan Ng, Tom Constant, Tom Wagstaff, Ulas Firat Tüzün, Verena Schuster, Victor Martinez, Wiem Borchani, William Aguilar, Xiaoxiao Ma, Yu Luo, Yuanliu Wanghan, and Zsolt Kegyes-Brassai.

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

Challenge and Data Providers: City of Bristol, City of Cascais, OpenWeather, City of Porto, Porto Digital, UN Studio, Urbanalytica, and Urban AI.

Financial Support: Capgemini, basecone, EDP, JTA The Data Scientists, SaltPay, Axians, DareData, Fidelidade, LTP, NOS, novobanco, Smartwatt, and Sonae MC.

Institutional Partner: EuroCities.

Jury Members: Aarthi Kumar, Ana Freitas, Bea Hernández, Carlos Gomes, Catarina Belém, Chanukya Patnaik, Daniel Ribeiro, Duarte Gomes, Erum Afzal, Filipa Castro, Filipa Peleja, Filipe Miranda, Floris Goes, Gabriela Lewenfus, Gilberto Titericz, Giovanna Miritello, Helder Filipe Oliveira, Henk van Dyk, Horácio Neri, Igor Quintanilha, Isabel Preto, Jacek Kustra, Jéssica Delmoral, João Vinagre, Joinal Ahmed, Jonathan Tooley, Julián Darío, Konrad Banachewicz, Kyra Wullfert, Márcio Rebelo, Mariana Rafaela Oliveira, Martina Pugliese, Nenad Tomasev, Nishrin Kachwala, Paballo Moeletsi, Paulo Maia, Pedro Lopes, Pedro Tourais Pereira, Ricardo Araújo, Rita Ribeiro, Rosana Gomes, Sudarshan Gopaladesikan, Telmo Felgueira, Teresa Scholz, Tiago Otto, Tom De Smedt, and Wenche Wang.

Finals Jury Members: Becky Belfin, Gitty Korsuize, Hubert Beroche, João Neves, and Mário Figueiredo.

Team Mentors: Alina Petukhova, Bernardo Monechi, Carolina Cerqueira, Catarina Freitas, Christina Caljé, Clarisse Magarreiro, Daniel Rodrigues, Daniela Costa, Elisabeth Fernandes, Emanuele Semeraro, Erum Afzal, Evan Simpson, Floris Goes, Francisco Amorim, Gabriela Lewenfus, Gonçalo Almeida, Henk van Dyk, Hugo Rações, Igor Quintanilha, Inês Gomes, Irina Vidal Migallón, Ivo Bernardo, Jennifer Gaskins, Joana Pereira, João Ascensão, João Gante, João Lobo, Joinal Ahmed, Jonathan Tooley, José Moreno, Lisa J. Knoll, Luís Espírito Santo, Maedeh Afshari, Mahmoud Abdel Aziz, Maik Reder, Matthias Kempe, Milton Santos, Nelson Nunes, Nikolai Janakiev, Nikolaos Lamprou, Nuno Moniz, Paballo Moeletsi, Paulo Maia, Pawel Potrykus, Pedro Chaves, Pedro Lopes, Ricardo Vitorino, Yasin Musa Ayami, and Yusuke Kaji.

Executive Summary

The main mission of the World Data League (WDL) is to create a hub of open knowledge on data for social impact. We do so by organising a data competition that brings data scientists from all over the globe together to solve socially-oriented challenges, focused on the United Nations (UN) Sustainable Development Goals (SDG). In 2022, the second edition of WDL took place with **7 different challenges** around the topic of **Data-Driven Communities**, focusing specifically on SDG 11 (Sustainable Cities and Communities) and SDG 15 (Life on Earth).

The competition's main outcome is open-sourced proof-of-concept algorithms that can help develop sustainable communities. The evaluation process is optimized not only for the technical evaluation but also for understanding the problem, analyzing the datasets, identifying possible applications of the algorithms in the organizations' day-to-day and measuring the social impact they will create.

This year, the novel framework for Social Impact Measurement was developed by the WDL team and introduced into the competition. The goal of this framework is to help define the product around the model being developed, the resulting outcome of the product, and the metrics that can measure the outcome together with their estimations.

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

Data describes the datasets the teams worked with, which were provided by real-world institutions. 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 submission. 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 used for technical implementation.

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 algorithm developed by the teams could be used and who would use it. This section aims to showcase possible products that could stem from these algorithms and identify the features, users, and outputs of those products.

Social Impact analyses the potential impact if the algorithm or product was implemented by the organization. This section aims to describe the desired social outcome and impact metrics.

The authors conclude that the second edition of WDL was a success, with many interesting outcomes and models that cities or organizations working with cities can benefit from.

Glossary

- **ALAN** - Artificial Light at Night
- **ARIMA** - Autoregressive integrated moving average
- **CMS** - Central Management System
- **EDA** - Exploratory Data Analysis
- **GTFS** - General Transit Feed Specification
- **HDBSCAN** - Hierarchical Density-Based Spatial Clustering of Applications with Noise
- **LSTM** - Long short-term Memory
- **OD** - Origin-Destination
- **OLS** - Ordinary Least Squares
- **POI** - Point of Interest
- **RNN** - Recurrent Neural Networks
- **SARIMA** - Seasonal Autoregressive Integrated Moving Average
- **WDL** - World Data League

Index

Introduction

Currently, data is generated in abundant amounts, to such an extent that it is hard for a human on his own to make sense of it all. Luckily, this abundance of data also sparked research into techniques and methodologies to interpret and even predict future outcomes based on this data. And thus, the profession of data scientist was born - a cross between software engineering and mathematical modeling.

These techniques have brought great leverage and advantage to corporations that knew how to use them but, being such a recent field, many sectors are still lacking behind in terms of knowledge and know-how. World Data League (WDL) aims to close this gap with organizations that are working on socially-oriented challenges. For this, we connected our challenges to the **United Nations Sustainable Development Goals**.

In 2022, we held our second edition with the topic **Data-Driven Communities** with over **200 participants** from **42 countries** that worked on 7 different **challenges over more than 4 months**. It was a very intense endeavor that produced **70 technical reports** responding to the proposed challenges.

How to interpret this document: This document aims at summarizing the used methodologies, showing the conclusions about datasets to solve specific challenges and presenting the main insights found by the teams. The authors would like to stress that the outcomes presented here should be considered as 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 that it is a direct or indirect cause. The results presented here are a summary of what the teams have presented in their reports. We hope that these ideas can spark future research directions, considerations for the data collected by cities to solve certain challenges and bring new ideas on how data can be leveraged to create social impact. All the ideas presented here can be found in the team's full submissions on the World Data League code repository.

Framework for Social Impact Measurement

In order to achieve our mission of creating **fast and accessible data-driven social impact solutions**, it becomes crucial to guarantee that all the solutions created by the participating teams take that into consideration. We believe that social impact measurement cannot simply be a bonus or a nice-to-have for the teams when developing this type of work, but instead **it needs to be an integral part of their work** and something they must bear in mind when submitting their solutions.

For that reason, we included social impact measurement of the solutions as a mandatory step and on which teams would be specifically evaluated. There is no point in creating an extremely powerful solution 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 submitting. This framework was the following:

1. DEFINE THE PRODUCT

- Define the input to the product and who gives that input (the "customers");
- Define the activity of the product and its features;
- Define the output of the product, what it shows to "customers", and the way it shows it.

2. DEFINE THE OUTCOME

- Define the long-term results, assuming the outputs are the immediate ones;
- Define what the product is supposed to achieve;
- Define the ''good'' that is being created.

3. DEFINE THE IMPACT METRICS

- Define 2 to 4 metrics that measure if the outcome is being achieved.

4. ESTIMATE HOW MUCH THE PRODUCT WILL IMPACT THE METRICS

- Since it is impossible to immediately see the impact of the product, estimate it quantitatively - for example, using the estimations/predictions of the model, market research, proxy industry products, or location.

WDL Topic: Data-Driven Communities

According to the United Nations (UN), 68% of the world population is projected to live in urban areas by 2050. With ever-growing cities, new challenges arise associated with population growth, but also a lot of interesting potential solutions. With the rise of smart city technologies, sensors, and open data initiatives, a data-driven approach is possible to develop those solutions.

In the last edition, WDL focused on the **11th UN Sustainable Goal: Sustainable Cities and Communities**. This year, we keep that core but enlarge our scope to embrace other challenges that highly condition our way to safer and happier communities. Our teams will have the opportunity to open new doors in subjects like climate, inequalities, and the environmental ecosystem.

The competition was divided into four different topics:

Environment

Transportation & Mobility

Safety

Biodiversity

STAGE1

Environment

Predict Waste Production for its Reduction

CHALLENGE BY
URBAN AI

According to the World Bank [1], in 2016 cities generated 2.01 billion tons of solid waste, which corresponds to 0.74 kg/day per person. With the rapid growth of cities, this number is only expected to increase and thus it becomes urgent to create optimization processes for waste processing and more targeted public education on waste management and separation. Finally, it is also important to note that waste collection also has a significant impact on air pollution [2].

The City of Austin is committed to a zero waste goal to reduce the amount of trash sent to landfills by 90% by the year 2040 [3]. Zero waste is a philosophy that goes beyond recycling, to focus first on reducing trash and reusing products and then recycling and composting the rest.

GOAL:

The goal of this challenge was to help identify trends in waste production and help to create insights into how to reduce waste and optimize its collection.

UNITED NATIONS SDG:

DATASETS AND PROVIDERS:

Daily waste collection data	City of Austin
Number of inhabitants per year	City of Austin
2020 Census data	City of Austin
Weather data	OpenWeather

DATA

Several teams resorted to Austin's open data portal to fetch additional data that could be useful for this challenge. Examples of such data are the waste collection routes, the recycling collection routes, socioeconomic vulnerability data, data about events and festivals in Austin and statistics about businesses in the United States, One Team mentioned that having access to datasets with metrics about past and current public policies may have helped correlate policies to socio economic cluster factors and waste trends already found. Another team found that names of roads were extremely inconsistent across open datasets, most likely due to renaming or merging of different roads.

METHODS AND TECHNIQUES

Data pre-processing mainly revolved around basic data cleaning - removing outliers and impossible values such as garbage collection numbers above the maximum capacity of trucks or negative ones. Another team looked into harmonizing the unit for analysis, which involved getting it to the census tract level by finding the attribution of waste of each census tract. looked into harmonizing the unit for analysis, which involved getting it to the census tract level by finding the attribution of waste of each census tract.

Regarding modeling, most teams used Prophet as a model for time series forecasting. Another team opted to use time embedding to produce weekly lagged data and use it to feed a traditional Machine Learning model, such as an XGBoost.

One team also did cluster analysis using k-means to cluster census tracts by their socioeconomic attributes with the intention of generating cohorts of census tracts.

MAIN INSIGHTS FROM DATA

Several teams pointed out that garbage collection constituted the highest volume of generated waste and the overall trend was increasing. One team also found that events play a significant role in waste production, as is the case of the SxSW Festival in Austin, which takes place in mid-March, and Christmas and New Year in December, the two months with the biggest waste production. This same team hypothesized that the March peak could also be due to an apparent seasonal trend of yard trimmings, since that type of garbage has a big peak during that month.

Another team pointed out that recycling-to-garbage ratio has been stagnating in the past years, which combined with growing population and retail consumption forecasts could pose a challenge to Austin's zero waste goal. This same team also noticed that in 2020 there were over 5000 cases when the same garbage route had to be re-visited outside the normal service day, which could be due to suboptimal truck allocation - their solution focused on solving this problem.

Another team who performed cluster analysis to the socioeconomic data, found that the clusters also had a strong geographical correlation and that they represented indeed distinct populations. For example, one of the clusters represented poor underdeveloped and minority prevalent census tracts - which also happened to be spatially close to each other. With this socioeconomic perspective and the waste forecast models, both on the same per census tract level, this team found that there was a possible improvement in waste recycling of 17.5M lb per month, if all census tracts reached the same recycling percentage of the best

representative for their respective cluster. They also plotted a quadrant view showing each cluster normalized by their population size in terms of how much waste they produced and how much waste they recycled.

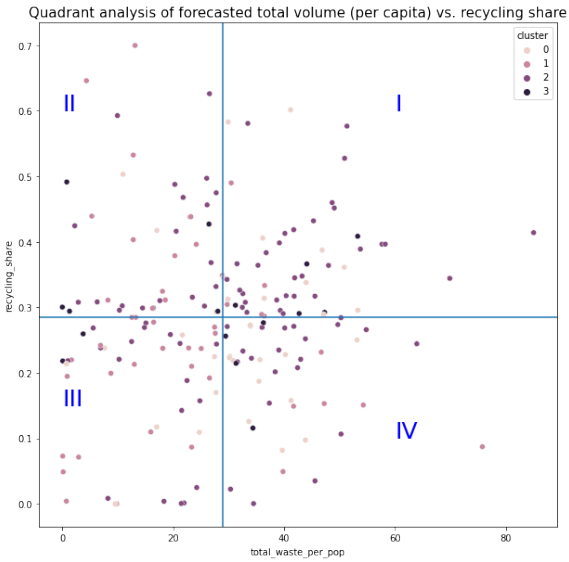


FIGURE 1

The four different behaviors in terms of total waste and recycling share for each normalized cluster.

- I - High recycling share and high total waste - desirable behavior where higher waste production rates generate high recycling rates.
- II - High recycling share and low waste production. The most desirable behavior.
- III - Low recycling share and low waste production.
- IV - Low recycling and high waste production are located. The least desirable behavior.

Plotting these four quadrants on a map helps understand which regions of Austin have a tendency towards certain types of behavior.

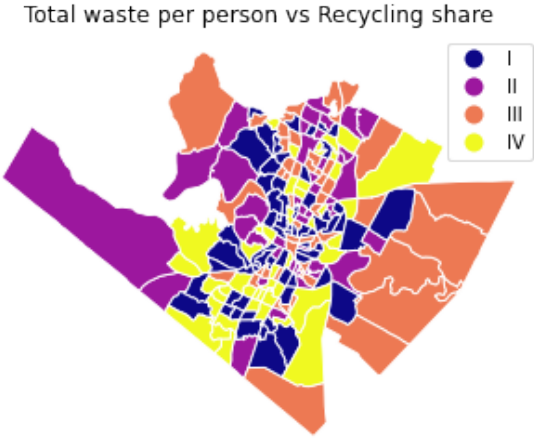


FIGURE 2

Map of Austin showing the regions color-coded by their type of behavior regarding waste production and recycling.

Using the forecast model, the team found that, in the current situation, 2022 would have a total recycling weight projected at 7.5 million lb per month. However, if the census tracts followed the behavior of their reference census tracts, an additional monthly 9.63 million lb could be recycled, decreasing the waste sent to landfills.

PRODUCT

One team proposed the development of an application that assists in the waste collection by planning collection trips along different routes based on the predictions for each route, whose primary user would be waste collection facilities. The application would suggest when to dispatch a collection truck on a specific route and for a particular load type based on the threshold values of the forecasting models. After each waste collection trip, the load weight could serve as a feedback input to the application to dynamically improve the schedules for the rest of the year.

Other teams suggested an Intelligent Decision Support System for policy decision-making regarding waste in Austin, whose primary users would be policymakers. This system would map waste generation in different regions and forecast waste and recycling per region within a tactical/strategic time. It could also generate cohorts of regions based on socioeconomic factors and provide macro-level target metrics based on the performance of the regions. This would enable the system to identify problematic areas due to the rapid increase in waste generation and low recycling performance.

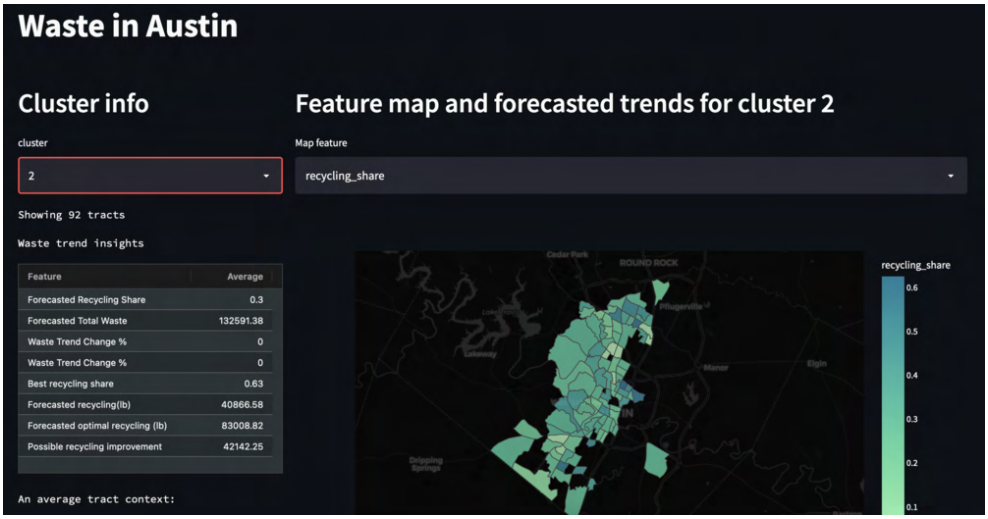


FIGURE 3

An example of a system that monitors waste production and recycling per region, showing macro-level target metrics.

SOCIAL IMPACT

One outcome of the proposed products would be monitoring and allocating city resources more efficiently, such as allocating garbage collection trucks.

One team suggested different metrics to measure such outcome: number of collection trips in a year, costs saved through better planning of trips, average capacity utilization of trucks, number of trucks added to the fleet per year and number of new waste separation and recycling facilities. Another team suggested measuring the number of extra collection days saved (i.e., no need to resend trucks because planned numbers were not enough to transport quantity) and variation of surplus in allocated trucks (i.e., more trucks allocated to a given area than necessary).

There is potential to reduce approximately 6000 trips a year for garbage collection and recycling single stream. Similar optimization studies have shown huge potential for savings for the civic authorities in addition to the qualitative impacts of less traffic disruption, less vehicle driver fatigue, and less pollution.

Another outcome of these products would be better planning of city policies by making data-driven decisions and implementing educational campaigns that improve recycling efforts by the local population.

As a way to measure this outcome, one team proposed evaluating the curbside recycling share and curbside total waste share across census tracts and the trend percentage change of recycling share and total waste.By connecting the best representative of all the city's regions with its socioeconomic descriptors and applying similar waste strategies to other regions with similar parameters, the team estimates a 9.63 million lb monthly reduction in incinerated waste. This solution would also translate into a considerable reduction of air pollution in Austin - 4368 tonnes of waste reduction lead to between 3057 and 7426 tonnes of CO_2 emissions [4].

Air Quality Prediction in Busy Streets

CHALLENGE BY UNSTUDIO

It is estimated that 9 out of 10 people worldwide live in places where air quality exceeds WHO guideline limits [5]. Due to high levels of air pollution, people risk getting diseases like respiratory infections, lung cancer, and heart disease. The most health-harmful pollutants are PM2.5 particles that penetrate deep into lung passageways.

The Green Mile is a project initiated by UNStudio, Blendingbricks, Heineken, the Rijksmuseum, the Amsterdam University of Applied Sciences, and the Dutch National Bank. It aims to transform Stadhouderskade street in Amsterdam, which is currently the most polluted, busiest, and the street with the most traffic and pedestrian accidents in the city.
The main sources of pollution are road traffic and industry; for that reason, people report feeling the effects of the bad air quality when spending large amounts of time in Stadhouderskade. As such, people are expectedly not attracted to spending more time there than strictly necessary [6]. Death rates attributed to air quality pollution have decreased in the Netherlands between 1990 and 2014 (approximately 45%) but plateaued in 2014 [7], which brings a renewed need to protect the air quality, not only in Stadhouderskade but everywhere.

GOAL:

Help the initiators of the project create a case and buzz for the needed change in Stadhouderskade street and, more specifically, for the current impact it has on air pollution.

UNITED NATIONS SDG:

DATASETS AND PROVIDERS:

Hourly measurements of different air pollutants at Stadhouderskade	UNStudio
Weather data	OpenWeather
Open city data	City of Amsterdam

DATA

Since no team used additional data to solve this challenge, only the provided datasets were used.

Several teams pointed out the critical relation between air pollution and weather conditions and how intrinsically related these two variables are. Wind, for example, plays a big role in determining the travel patterns of air pollutants since it can transport them. For that reason, one team mentioned that having hourly measurements of air pollution but only daily weather measurements posed a problem in analyzing the data.

METHODS AND TECHNIQUES

All teams started with EDA by analyzing the descriptive statistics of each variable and their pairwise relations through scatter plots and correlation values. Besides that, all teams looked into variations across different time frames and possible missing data.

During data cleaning, one team established a maximum threshold for pollutant variables after identifying unusual/extreme values in the series using a moving average plot. This team fixed missing data problems using linear interpolation for missing observations that were, at most, one day apart from a known observation. The remaining missing values were discarded from the analysis. Another team used a 3-point rolling mean to fill null values.

Regarding feature engineering, several teams computed the Common Air Quality Index, which provides a unified view of the air quality at any given moment, taking into consideration three of the measured pollutants: Nitrogen Dioxide (NO2), Particulate Matter 2.5 (PM2.5), and Particulate Matter 10 (PM10). One team also calculated mutual information, permutation importance, and Principal Component Analysis. In terms of time series modeling, several teams evaluated stationarity and autocorrelation using the Dickey-Fuller test and used Autoregressive integrated moving average (ARIMA) or SARIMA (Seasonal Autoregressive Integrated Moving Average). Others used XGBoost and LightGBM, but performances were not significantly better.

MAIN INSIGHTS FROM DATA

Several teams discovered that all air pollutants showed a decreasing trend from 2014 to 2022, ranging from 11% to 81%. Compounds Xylene (81%) and Toluene (71%) decreased the most, which could point to the fact that Stadhouderskade was already on the right track to decreasing air pollution.

One team used geographical data related with the location of outdoor activities in the nearby zones of the Stadhouderskade street to show that there were no running routes in this street and that there was only one sports park in the vicinity of this road. This same team also showed there was strong traffic congestion since there was a high concentration of pollutants usually emitted by motor vehicles.

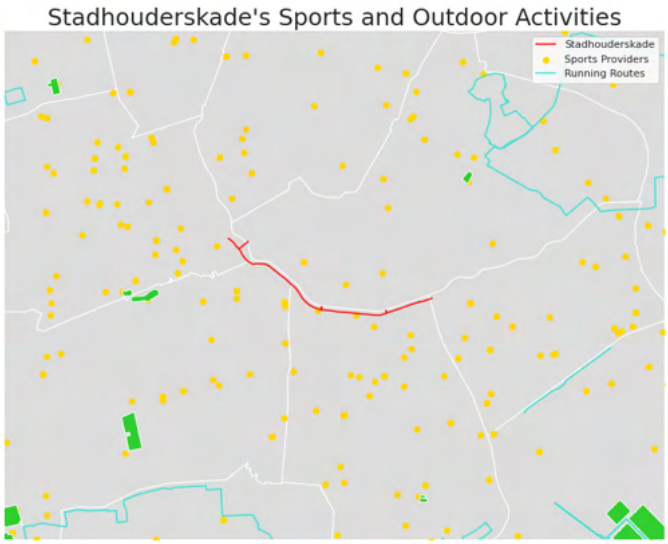


FIGURE 4

Map showing Stadhouderskade street (in red) and its surrounding infrastructure (in blue, yellow and green). There were no running routes on this street, and there was only one sports park in the vicinity of this road.

A team found that except for NO2 - whose values were lower in the early mornings when compared to the entire day - no other pollutants showed similar concentration patterns. However, there seems to be a pattern throughout the year: from May to August (Summer), the pollutant value decreases and the air quality index increases; in December, the pollution levels increase considerably.

PRODUCT

As a way to productize the developed algorithm, the vast majority of teams suggested developing some type of dashboard or application that would enable city planners to view the different levels of air pollutants at any given time, along with predictions for other times in the future.

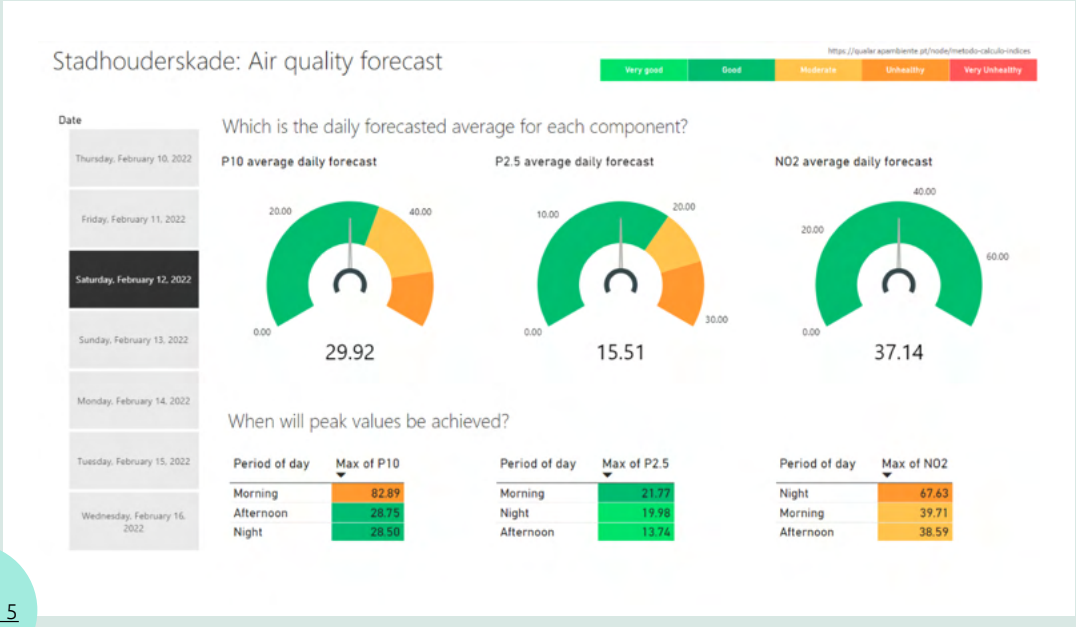


FIGURE 5

Dashboard showing the different level of each air pollutant on different days. The user can also view predictions of when the peak values will be achieved.

SOCIAL IMPACT

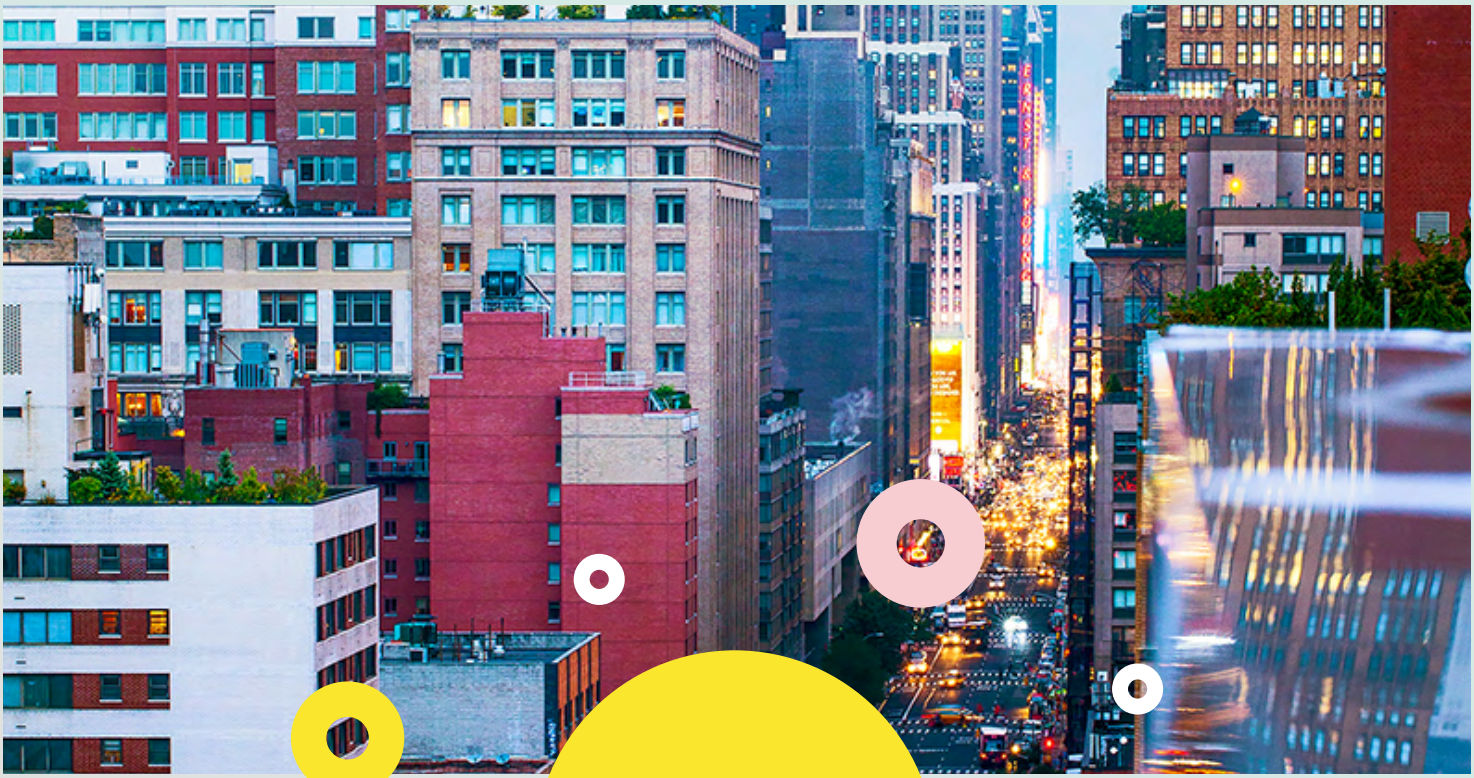
The main outcome of this product would be changing the city policy using the gathered data. One team suggested the following examples:

- Implement suggestions to the population if the Air Pollution Levels are "high" or "very high" - for example, suggest that people with respiratory diseases avoid passing the street at certain times of the day.
- Create a "Low Emission Zone" where only cars registered after 2011 can circulate in Stadhouderskade, in times of the day when Air Pollution Levels are "high" or "very high" - based on the example of Krakow (Poland), the implementation of car traffic restrictions could lead to an 80% user satisfaction towards the quality of the public space. [8]
- Optimize and add public green spaces in the vicinity of the street: fewer parking lots and more green areas, gardens, or parks.
- Build outdoor interactive banners along the street displaying the amount of air pollutants emitted over a certain period, where people could filter and visualize these amounts in real-time.

Some metrics to measure this outcome would be the average number of pedestrians passing on the street, the monthly percentage of "high" or "very high" Air Pollution Levels on the dashboard, and the monthly number of green space users.

Another team proposed as metrics the number of days with acceptable/unacceptable levels and the percentage of air pollution decrease after deployment of their product.

There was also a team proposing that by using their analysis, city planners could create efficient traffic control policies (i.e re-routing traffic in certain times of the day) or even create additional anti-pollution policies, like limiting the usage of specific fireworks in New Years to reduce the pollution levels in critical moments. On another note, the analysis could also be used to create articles or media campaigns to generate social conscience on the pollution problem.



STAGE2

Transportation & Mobility

Optimization of public transport routes during road interruptions

CHALLENGE BY
CASCAIS

Data from the United States of America shows that in 2019 alone, Americans took 9.9 billion trips on public transportation [9]. Research also showed that public transportation provides economic opportunities, is safer to travel than cars, saves money, reduces gasoline consumption, and reduces the carbon footprint.

On the other hand, cities are constantly being redesigned and maintained. Sometimes it is necessary to perform interventions on the public road and resort to traffic cuts or drifts. These disruptions on the public road cause inconvenience to the residents of the affected streets and the city's entire mobility system, including public road transport.

If these cut-offs/drifts are constant, this can create situations of distrust in the reliability of the public transport network. Thus, it is crucial to ensure that the re-routing of public road transport has a minimal impact on the users.

GOAL:

Model which routes of the transport road network suffer the most disturbance due to interventions on public roads, and to evaluate the efforts needed to adapt services to match the network's needs in the presence of such disruptions. Additionally, the goal was also to assess and quantify levels of perception of "inconvenience" by network users caused by different disruptions.

UNITED NATIONS SDG:

DATASETS AND PROVIDERS:

GTFS Public Transport Network	City of Cascais
Bus Routes	City of Cascais
Road Network	City of Cascais
Historical interventions in public roads	City of Cascais

DATA

The main limitation identified by the teams was the lack of ground truth for the disturbance. It would have been helpful to know the scheduled time of arrival at each bus stop and the real arrival time. Another aspect was the location of the interventions; while the dataset included the street name, it would have also been useful to include the GPS coordinates. Besides this data, one team also included the POI's extracted from Google Maps for the analysis.

METHODS AND TECHNIQUES

Three different methods were employed to approach the optimization problem in case of an interruption. The most straightforward approach was the creation of a distance matrix between several stops and finding the nearest stop for the alternative stop.

The rest of the teams focused on a graph approach. More specifically, the Dijsktra's Exact Algorithm was used to find the shortest distance in case of an interruption. The teams measured the inconvenience through either the extra time needed to travel to the original stop or the wait time.

MAIN INSIGHTS FROM DATA

It was found that the road interruptions were unevenly geographically distributed and that some areas had a higher number of interventions, as seen in Figure 6. The teams also noted that the highest reason for interruptions was due to work on the water supply.

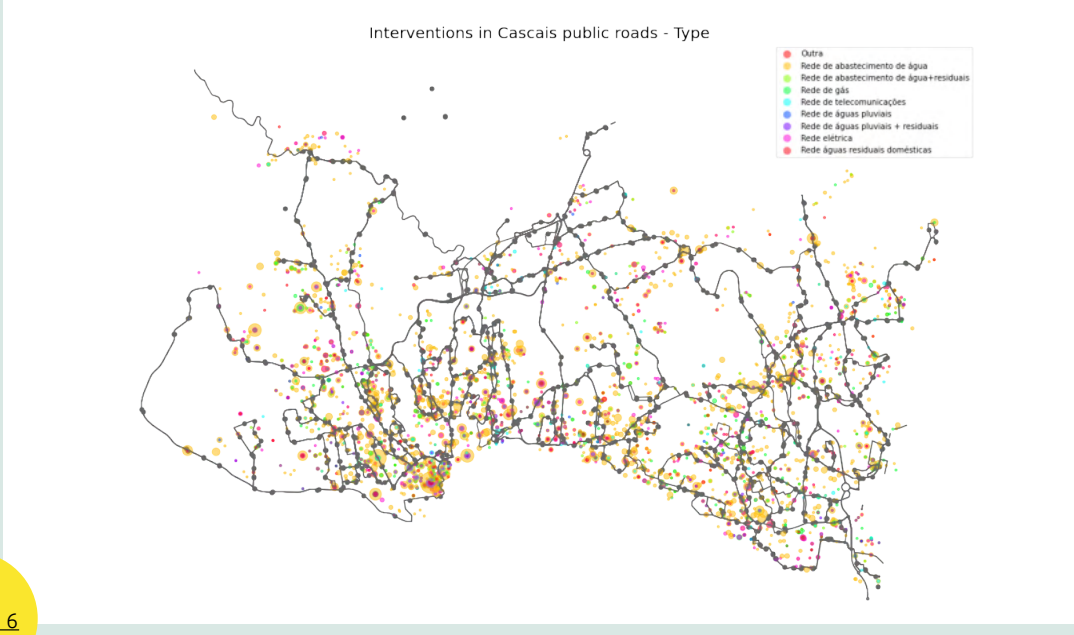


FIGURE 6

Map showing the different types of road interruptions. The most significant number of interruptions was due to water network works represented in orange.

PRODUCT

The proposal for products in which to implement the above models revolved around creating a dashboard for city officials to plan the road interventions and the detours for public transport. The dashboard would show how severe road work impacts users and suggest roads for re-routing the buses.

As a side product, some teams also suggested integrating into the Cascais mobile application the possibility to send push notifications with planned changes to the route.

SOCIAL IMPACT

The primary outcome would be the possibility to predict the inconvenience caused by the planned road interruptions, informed users, and a better temporal distribution of road interruptions. The suggested impact metrics were the reduction in wait time during road interruptions and in the walking distance to original bus stop changes. It was also suggested to measure qualitative feedback from users on the app.

Predicting the flow of people for public transportation improvements

CHALLENGE BY

Porto.

ASSOCIAÇÃO PORTO DIGITAL

As cities grow, the flux of people moving from and to cities increases. While many people travel by public transportar, some do not due to service levels that might not be acceptable. [10] Nonetheless, according to TomTom's traffic index [11], if a person were to drive every day during rush hours (typical commute time) instead of non-rush hours, that person would spend an extra four whole days inside of their car. On the other hand, it was found that compared to cars, public transportation produces 95% less carbon dioxide and 92% fewer volatile organic compounds. [12]

Therefore, by increasing public transportation usage, it is possible to decrease pollution, improve public health and increase the general well-being of the population. While many factors affect public transportation usage, the quality of coverage is a crucial. With access to mobility and public transportation data, there is an opportunity to optimize the public transportation system.

GOAL:

Study how the city's public infrastructure can be improved to help reduce traffic and improve the quality of life for its citizens. For this, it was asked to create a model that predicts the in and outflow of people short-term (days) and long-term (months) to and from the municipality of Porto.

UNITED NATIONS SDG:

DATASETS AND PROVIDERS:

Entry and Exit validation data from public transportation in the Metropolitan Area of Porto	Associação Porto Digital
Origin-Destination (OD) matrices of Movement of People from/to the Porto Metropolitan Area	Associação Porto Digital
GTFS from Porto's Metro and Public Bus System	Associação Porto Digital

DATA

The teams identified data quality issues and highlighted the importance of clean datasets for increasing the model's performance. The quality issues identified were missing values or values that were clearly out of distribution. On the other hand, it would have been interesting to have a better spatial resolution in regard to the mobility data. This is because the origin-destination matrix was only at the municipality level, which is a vast area to optimize.

Besides the provided data, the teams included data regarding big events in the city, holidays, COVID restrictions, and weather conditions.

METHODS AND TECHNIQUES

The teams used different approaches to predict the flow of people to and from the city. Some focused on more classical time-series approach such as ARIMA, while one team did an extensive analysis of the seasonality and stationarity using the Dickey-Fuller test. Other teams focused on deep learning techniques such as autoregressive Recurrent Neural Networks (RNN and Long Short-Term Memory (LSTM).

MAIN INSIGHTS FROM DATA

All teams found that there is seasonality in the data with well-defined peak hours for each day and that the commutes are reasonably predictable. As expected, there is a large flow during the working days, and a decrease during weekends, with peaks of flow near 8 AM and 6 PM on weekdays.

One team noted that some of the public transportation was not optimized - that is, while the rate of ticket validations varies, in many cases the rate of available transport does not. An example of this effect can be seen in Figure 7.

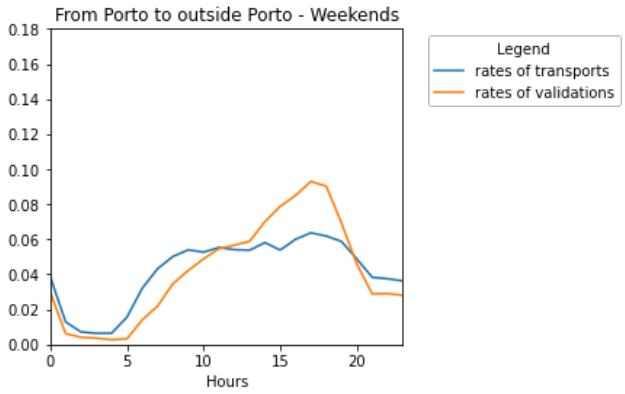


FIGURE 7

An example of a misalignment between the rate of ticket validation (orange) and the rate of transportation (blue).

PRODUCT

Most of the suggested products targeted the person responsible for planning the city's public transport system and schedules. The proposed product was a dashboard that can predict the flow of people between different stations and locations to optimize the scheduling in the upcoming season. One approach was specifically directed to the Metro of Porto for predicting the short-term affluence for the increase or decrease of frequency/number of carriages.

One team suggested creating an app for bus users offering an incentive for commuters to travel in off-peaks which would be predicted by the model. On the other hand, it could warn regarding a possible increase in usage, and therefore inform the user about their commute peak and off-peak hours. Both suggestions aim to improve service quality and increase the number of users.

SOCIAL IMPACT

The teams predicted that implementing the above products could potentially improve the quality of service of public transportation and therefore increase the number of users and reduce the use of cars. Several metrics were proposed to measure the impact:

- Number of passengers using the public transport
- Number of cars circulating in the city
- Size of the traffic jams
- Growth of the usage of public transport
- Level of satisfaction of the users
- Level of air quality in the city

While many metrics were proposed, no team estimated the improvement of the metrics in the case of implementation of the product.

Optimization of soft-mobility drop-off points

CHALLENGE BY

Porto.

ASSOCIAÇÃO PORTO DIGITAL

A recent review [13] found that soft mobility (including e-scooters), depending on the city and culture, is used differently for entertainment purposes or commuting. Specifically for commuting, the usage of soft mobility has the potential to help with the last mile problem. This is also backed up by the same review, which shows most trips have a distance of 0.72–2.4 km and last, on average, between 8–12 minutes. On the other hand, according to the Tom Tom traffic index [14], Porto's traffic is worse than, for example, Madrid's traffic, where residents spend 41 hours per year in traffic jams in comparison to 52 hours in Porto. A good transportation system is crucial, and soft mobility might be vital in improving how citizens move around the city.

GOAL:

Study and analyze the soft mobility pattern in Porto to improve the overall experience and usability. More concretely, the teams were challenged to create an optimization model for optimizing the drop-off locations of the e-scooters.

UNITED NATIONS SDG:

DATASETS AND PROVIDERS:

E-Scooter Transport Data	Associação Porto Digital
E-Scooter Location Data	Associação Porto Digital
GTFS for E-Scooter Parking and Metro Stations	Associação Porto Digital
Entry and Exit validation data from public transportation in the Metropolitan Area of Porto	Associação Porto Digital
Origin-Destination (OD) matrices of Movement of People from/to the Porto Metropolitan Area	Associação Porto Digital

DATA

Most of the teams enriched the dataset provided by the City of Porto by downloading Points of Interests (POIs), the road network, schools, and city boundaries from OpenStreetMaps. One team used the taxi trajectories to evaluate if a specific place is a hotspot. Demographical data regarding each parish was also used.

METHODS AND TECHNIQUES

Most teams started by creating candidate points with two different approaches: selecting them manually (e.g., current drop-off points, bus stops, and metro stops) or automatically determining them through a clustering algorithm. For example, one team used HDBSCAN and others used k-means clustering.

All the teams, later on, applied an optimization algorithm as defined by the constraints given by the challenge provider. The optimization methodologies used were Constrain Integer Programming, PuLP Linear optimization and mixed integer programming.

MAIN INSIGHTS FROM DATA

In the data, the teams observed that the e-scooters mainly were used for recreation purposes and not as the last mile travel solution. This was seen in the data as having a more than double drop-off rate near recreational places rather than bus stops, while there were more bus stops than recreational points. This was also visible when compared to the days of the week and hours that e-scooters were used. In Figure 8, it is possible to see that the most use was at the end of the day and during weekends.

The teams did not see an expected peak during morning rush hour, and they also noted that there were many standard routes that people used (e.g., in the old town or by the sea), and this, in turn, could be used to improve safety for the most used routes.

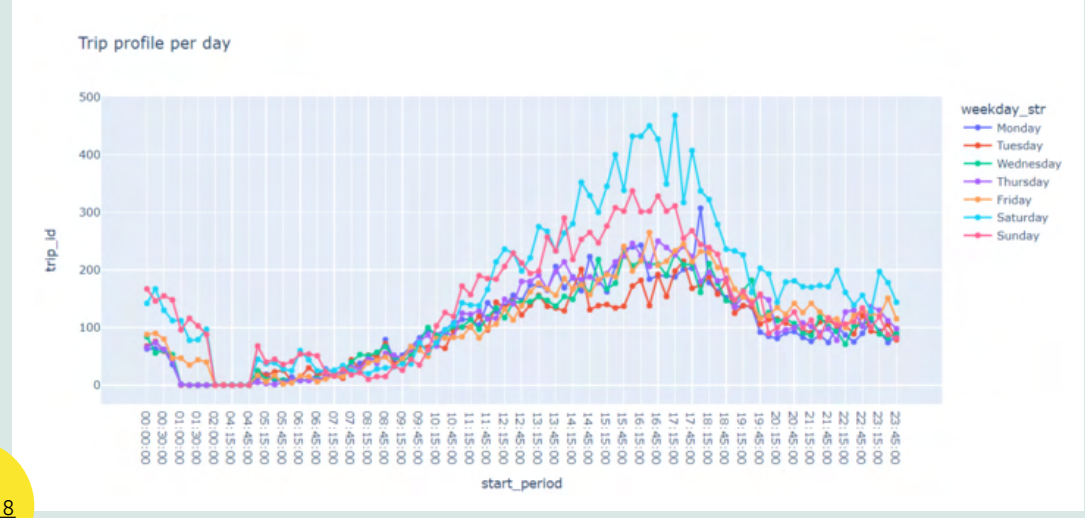


FIGURE 8

The number of trips is higher during the weekend than during the weekdays. This could be because the scooters are mostly used for entertainment purposes.

PRODUCT

All the teams proposed a similar product: a dashboard for service providers, government bodies, or regulatory agencies to optimize the drop-off zones and monitor their use. One team additionally proposed to create live information on the performance of each drop-off zone and indicate if it should be changed. Figure 9 shows an example of a dashboard where the user could regulate the weights (that is, the importance) for each factor to be considered when optimizing.

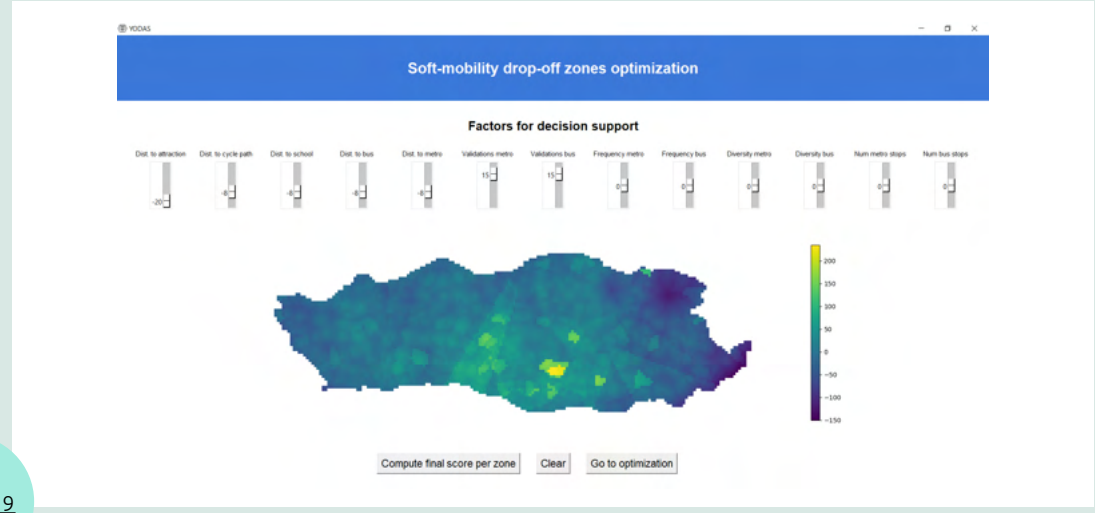


FIGURE 9

An example of a dashboard for optimizing the drop-off zones.

SOCIAL IMPACT

By creating an optimized e-scooter network, the e-scooters should be easier to find, more useful to citizens, and constitute a better alternative to cars. As metrics to measure social impact, the teams proposed:

- Total number of scooter trips
- Number of scooter journeys per day and increase compared to baseline
- Percentage of scooters correctly dropped in the drop-off zones
- Number of cars circulating in the city
- Maintenance cost
- Lifespan of e-scooters
- Pickups and dropoffs: the number of e-scooters needed to be relocated

Teams estimated an increase in the number of scooter journeys between 2.53% and 12%. One team estimated this expansion based on their model and another team made the estimation based on previous successful cases. The same team also calculated that there was a potential for a reduction in CO2 emission per transportation mile of 80.2 tons.

SEMI-FINALS

Safety

Predicting a safety score for women in Costa Rica

CHALLENGE BY
Urbanalytica
COLLABORATIVE PLATFORM OF URBAN INNOVATION

According to a study by the United Nations Entity for Gender Equality and the Empowerment of Women [15], gender violence in cities, specifically in public spaces, has become an increasingly public issue, especially in Latin America. The lack of adequate urban infrastructure, policies, and governance models exacerbates it. Thus, addressing the main obstacles women face regarding their right to an inclusive and safe city becomes a priority.

Police statistics have shown that 70.6% of the complaints of street sexual harassment in Costa Rica in 2019 were submitted by women [16]. While no current strategy from the public authorities is in place, women are raising their voices and creating awareness groups on social media, for example, to report aggressions and missing persons. This is why women need a mapping tool to identify and report whenever they feel their right to enjoy public spaces without being harassed is being threatened.

Tools like this already exist in other parts of the world. For example, in India, the Red Dot Foundation created the SafeCity web app.

GOAL:

Create a safety index to assess conditions, insecurity, and gender violence in public spaces affecting women and girls in Costa Rica and predict its trend.

UNITED NATIONS SDG:

DATASETS AND PROVIDERS:

Police reports from 2010 to 2022, including the type of crime, location, and anonymized information about the victim	Urbanalytica
Information about the location of POI, commercial and residential areas, road network, and other public infrastructure	Urbanalytica
Demographic and crime rate data	Urbanalytica
Google review data	Google
Weather data	OpenWeather
Open city data	National Statistics Institute of Costa Rica

DATA

While many teams recognized the richness of the datasets in terms of time span, the lack of geographical granularity was noted as one of the weak points. One team simulated how a sample of such a dataset would look like.

Another team used the Penal Code of Costa Rica as an additional source of data regarding the severity of the crimes - the higher the jail time, the bigger weight that crime would have on the index.

There was also a team enriched the dataset by conducting a survey to young adults of their origin country that was centered on people's perception of safety, for which they obtained 153 responses. This same team also noted that it would be interesting to have additional data on the flow of people between different points of the city and more demographic data.

A different approach was gathering additional data from OpenStreetMaps regarding public street lighting, as a way to measure the correlation between public lighting and crimes committed.

METHODS AND TECHNIQUES

All teams started with some level of exploratory data analysis, mainly around the distribution of each type and subtype of crime, along with its prevalence per age, gender, location, and different time intervals..

One team started by forecasting gender-based crime. They considered a crime gender-based if its prevalence per gender was above the average proportion plus half of the standard deviation. This team then analyzed the autocorrelation and seasonality of the data and trained a forecasting model to predict the number of crimes using Facebook's Prophet algorithm - which, compared to a naive model baseline, performed 10% better.

After that, the team moved on to compute a risk score considering the following variables as positive indicators of safety: lighting, width, and type of roads, visibility, population density, presence of security facilities, public transportation, and diversity of people. They then calculated the safety score for each polygon on a map and merged that with the crime data for that same polygon in an effort to find which variables had more influence on crime using an Ordinary Least Squares (OLS) algorithm. Most features had statistical significance in the model, and the ones that did not were removed. Finally, they checked spatial autocorrelation between neighboring polygons using Queen Contiguity - because hexagons could have up to six neighbors each - and found positive spatial autocorrelation in the number of crimes with statistical significance, meaning they were clustered among neighbors.

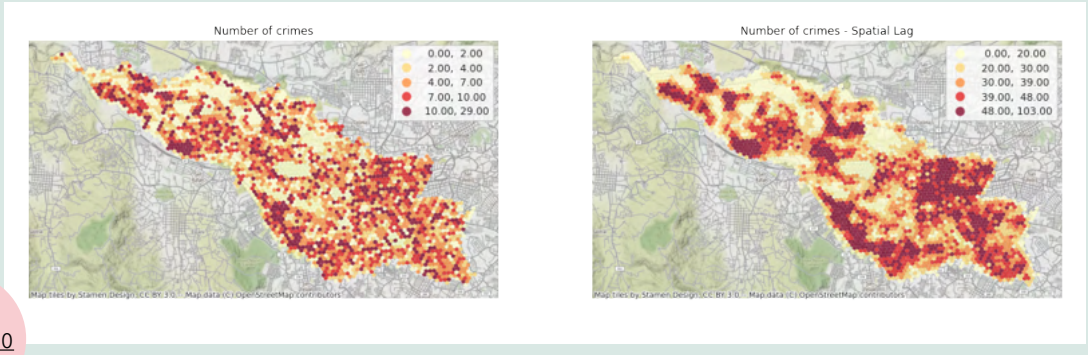


FIGURE 10

Maps showing the distribution of the number of crimes (left) and the spatial lag of the variable (right). Spatial lag explains the influence of the neighbors in the data. There are patterns in the data, clustering the crimes in certain zones of San Jose.

Another team calculated their safety risk by simply looking at the crime data per district of San Jose. They studied the Costa Rica Penal Code to find the number of sentence years for each crime as a way to differentiate between the severity of crimes and quantify it. Additionally, if the crime was committed against a minor or against a woman, its severity was multiplied by 2 - this methodology was based on the Pinkerton Crime Index [17].

Finally, they took the sum of the incidence index for each district at each quarter and divided this number by the population of each district at that point in time, so that the index captured the frequency and severity of incidents and the size of the population in that area. For prediction, using pre-2019 data to train and 2019 data to test, this team trained a Linear Regression model, with a Mean Squared Error of 0.48 for yearly predictions and an XGBoost model, with a Mean Squared Error of 0.61 for yearly predictions.

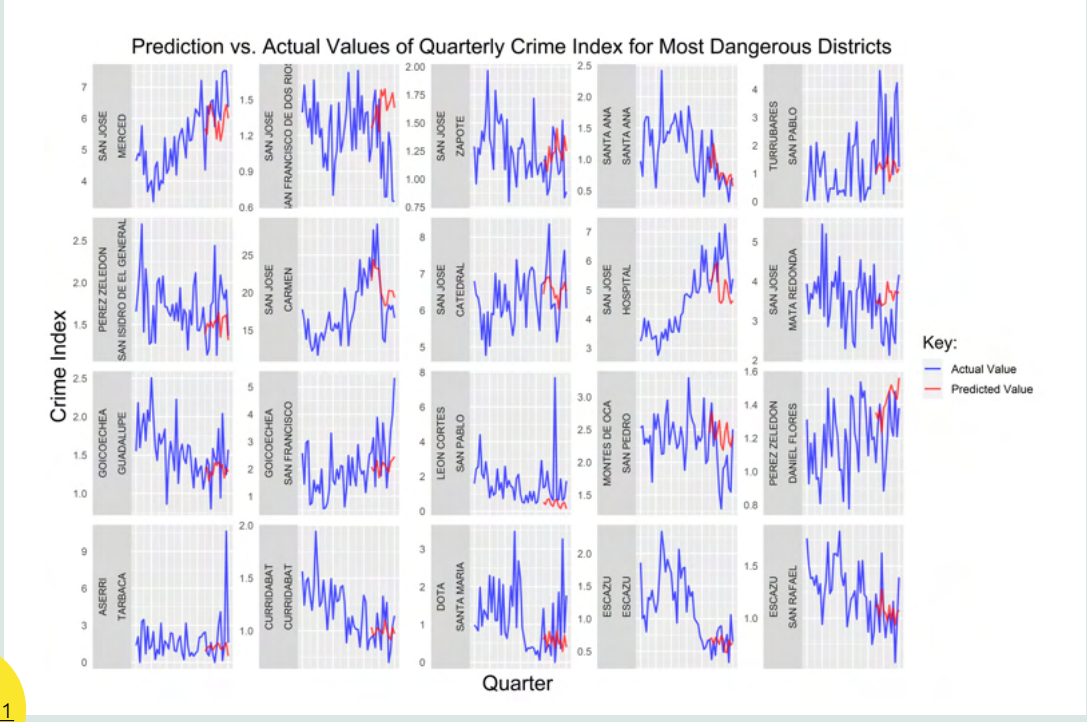


FIGURE 11

Quarterly prediction of crime index on the top-20 worst districts for each quarter, using the Linear Regression model.

Another team, while not approaching the problem in a radically different way, did provide a different product solution that lead them to develop solutions in route optimization. This team used a Dijkstra's Shortest Path algorithm to find zone-based paths between two points that minimized the traveled distance and their overall safety index. The rationale behind the algorithm was to create a cost matrix depicting the costs between adjacent zones: distance and score.

MAIN INSIGHTS FROM DATA

Several teams found that the most prevalent crimes were theft, assault, and robbery, with homicides being extremely rare. They also found no difference across months or days of the week but found that most crimes occur at night. In terms of gender-based crimes, the most prevalent crimes targeting females more than males were outbursts (64%), femicide (96%), and domestic violence (59%).

One team found a strong correlation between the presence of road infrastructure, properties, and heritage buildings and the increase of crimes. On the other hand, recreational areas, institutional, commercial, mixed land use, and population increase the safety of the space. They also found that the neighboring surroundings have an important influence on the number of crimes in a certain area, meaning crime has to be tackled holistically and not only street by street or block by block.

Another team created a very simple and easily interpretable safety index based solely on crime data - explained in the section above. With this index, they found that although women were victims in only 35% of crimes, proportionally to the crime severity captured by their index, that number climbed to 50%. The same thing happened in the distribution by type of crime - for example, while assault represented 38% of crimes and robbery only 14%, after adjusting for severity, those proportions became 21% and 35%, respectively. Another interesting finding is the variance across districts and years. For instance, this team found that in the district of Carmen, in 2010, the crime index was 16.1 - this means that Carmen was 16 times more dangerous than the San Jose average in the same year. However, Carmen was only the 5th district with more crime reports in 2010, which means that although fewer crime occurs there, per population and severity, their type of crime is considerably more serious.

PRODUCT

Most teams suggested productizing their algorithms by creating an application or website that displayed maps, statistics, and forecasts about gender-based crimes in Costa Rica, allowing future crime reporting and forecasting. The users would mostly be female inhabitants and authorities.

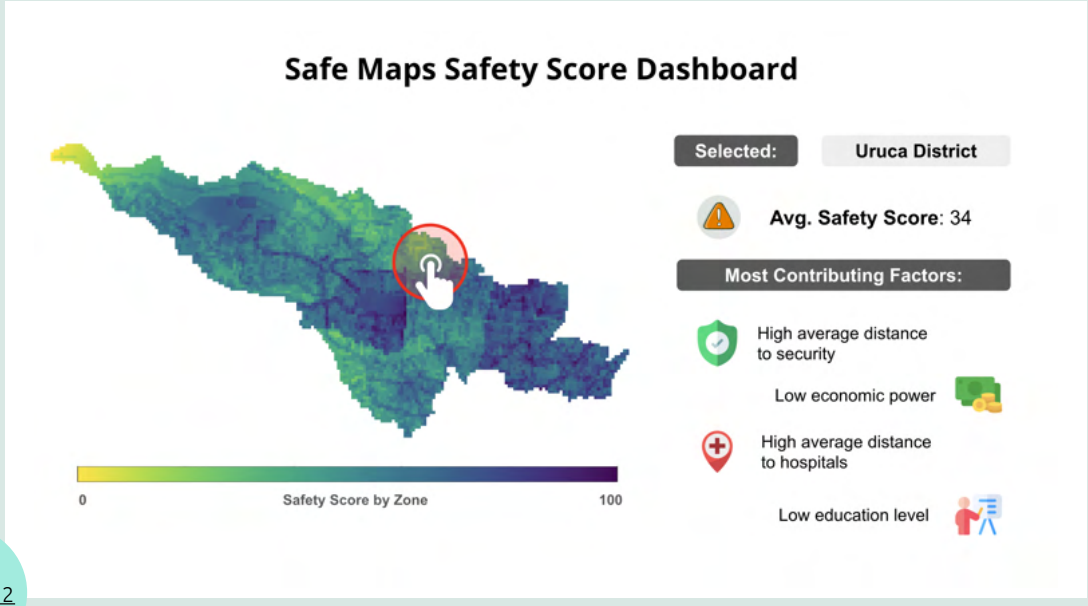


FIGURE 12

An example of a crime index map, color-coded by crime and safety index, together with contributing factors to explain that result.

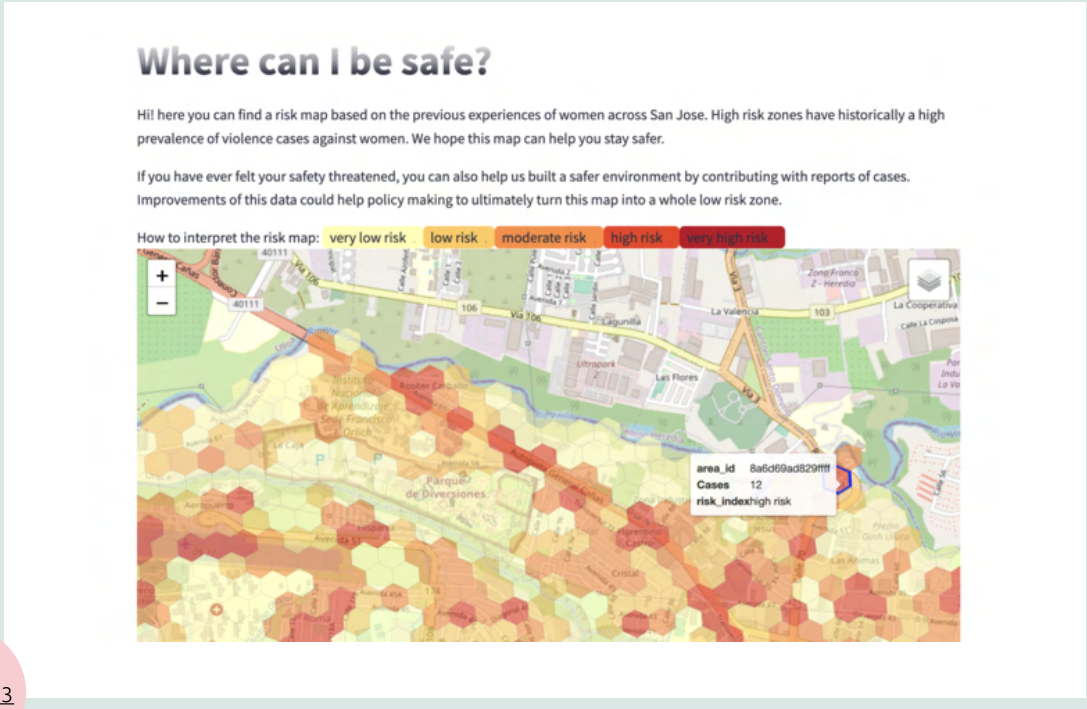


FIGURE 13

An example of a crime index map, where different areas are represented by hexagons and the safety index is represented by a color.

There was **a team** that suggested a navigation system that would suggest routes taking into consideration their crime and safety index, which could be used by women and tourists. This system would compute the shortest, safest, and "danger threshold" paths between different zones and enable users to choose the route they feel most comfortable with.

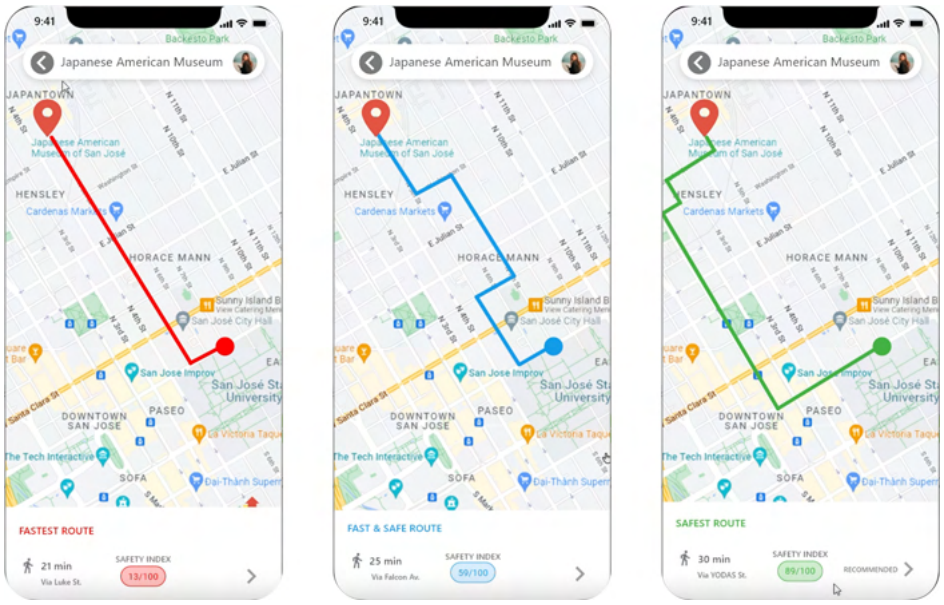


FIGURE 14

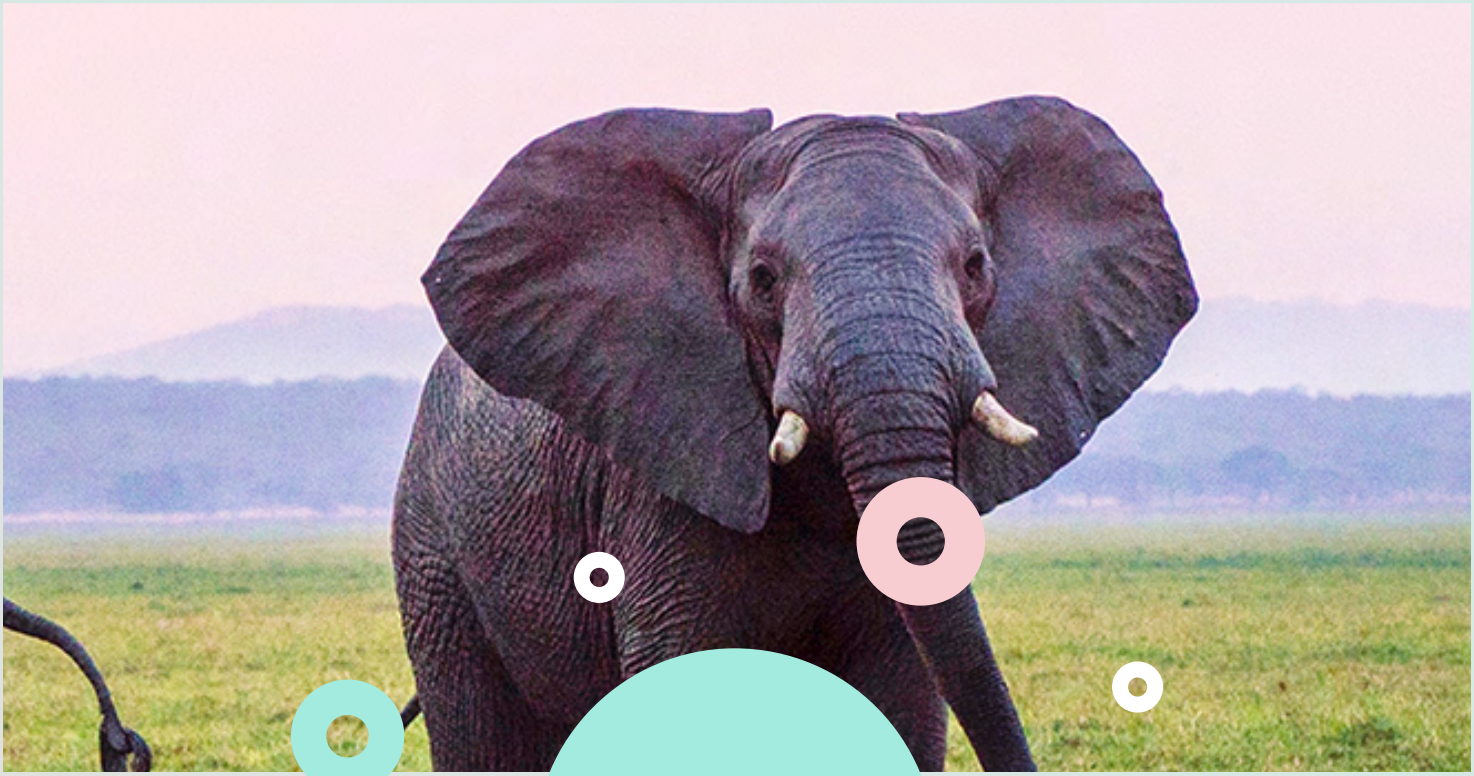
An example of a navigation system that suggests routes based on the crime and safety index.

SOCIAL IMPACT

The main outcome identified by all teams was the decrease of gender-based violence by reducing the number of street crimes, including sexual harassment. A secondary outcome would be increasing the amount and reliability of gender-based violence data by making it easier and more comfortable for women to report sexual harassment incidents.
Looking at the primary outcome, the main metrics to assess that would be the number of reports of gender-based crimes, the number of individuals persecuted because of committing gender-based crimes, and the number of areas with a decrease in gender-based crimes.

Based on model and survey predictions, one team estimated an increase in people's safety of 9.09% by having the option to choose safer paths.

Another team pointed out that several publications found evidence that urban planning and design improve the liveability of cities and towns. For instance, with a randomized experiment in New York City, evidence showed a 35% reduction in outdoor nighttime index crimes [18]. In the case of Costa Rica, according to the available data, more than 57% of crimes occurred at night, and although we cannot say whether outdoors or not, even in the case these were only 30% of the total, this product would result in a reduction in the crime index of 10%.



FINALS

Biodiversity

Identification of Dark Ecological Corridors

CHALLENGE BY

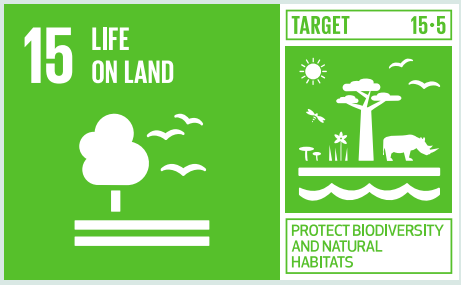
Artificial Light at Night (ALAN) is increasingly recognized as a major threat to global biodiversity [19]. It alters the amount, quality, and connectivity of available habitats for species. Light pollution causes habitat fragmentation by making areas harder to pass through and thus creating spatial barriers, evolutionary changes, and distorting normal growth patterns, among other negative effects. Lighting affects species differently depending on when their feeding and mating seasons occur.

One of the most affected species is horseshoe bats [20], as they are sensitive to light, and ALAN can greatly reduce their feeding area and activity. As a solution, dark ecological corridors and spaces can be created inside the city so bats can thrive and move around different areas. Public lighting systems today enable temporarily activating elements of the dark ecological network. The Bristol City Council (BCC) plans to install a Central Management System (CMS) within the next two years to allow dimming regimes to be implemented in certain locations.

GOAL:

Reduce the impact of ALAN on the bats in the city of Bristol and therefore reduce its adverse effect on the ecosystem. To accomplish this, the teams needed to create an optimization algorithm connecting the bats' natural habitats (feeding, roosting, and breeding sites).

UNITED NATIONS SDG:



DATASETS AND PROVIDERS:

Records of recently spotted bats (within the last 10 years) with 1km resolution for security purposes. Bat records older than 10 years were provided at full resolution	Bristol Regional Environmental Records Center (BRERC)
Data on the occurrence of moths (a major food source for bats) in Bristol	BRERC
West of England Habitat GIS Map with Priority Habitats, potential Priority habitats and other habitats	BRERC
Green Alleys GIS	BRERC
Wildlife Corridors dataset containing sites that help link up and buffer Sites of Nature Conservation Interest (SNCIs) and the City Green Belt	BCC
Ordinance survey open data green space	BCC
Public lighting data	BCC

DATA

Besides the data provided, most teams complemented the datasets with information on POIs and transportation services from OpenStreetMaps. Some datasets from the Bristol Open Data Portal, such as traffic accidents and criminality rates per ward, were also used. This data helped to understand where lightning might be more important compared to other places.

One team used satellite information provided by the *NASA Visible Infrared Imaging Radiometer Suite* that quantifies the amount of light reflected from Earth. Using this dataset, it was possible to understand the areas where it might not be possible to turn off the lights (e.g., billboards).

Regarding the data provided, most teams pointed to the quality of the observational datasets. While the data goes many years back, the observations were not consistent; for example, some years had only one datapoint measured. Although it made yearly modeling possible, it posed a challenge for seasonal modeling. Another issue found with the observational data was that categories such as "several", "present", and "abundant" had no numerical definition. It was also pointed out that making the crime dataset (available on the city's open data portal) more detailed regarding the location would help the model include more safety parameters.

METHODS AND TECHNIQUES

The teams divided the map into a grid (most used hexagons). Some teams started by using an HDBSCAN and X-means for clustering areas of bats or prays that needed to be connected. One team also clustered together the location of bats and prays and applied the Lotka-Volterra equations to describe the dynamics of each cluster.

Some teams used Dijkstra's algorithm to identify the dark corridors to find the shortest path considering the number of bats, moths, street lights, and penalties for switching the lights off. One team, on top the Dijkstra's algorithm, used Markov chain processes on graphs to mimic the bat's behavior and an Agent-Based Model to achieve an overall optimal street lighting configuration.

Another approach was to use a genetic algorithm to identify the dark ecological corridors.

MAIN INSIGHTS FROM DATA

It was noticed that the number of horseshoe bats in the last decade has been decreasing in the city of Bristol, therefore highlighting the importance of this work. One team also observed that horseshoe bats could not be compared to other bats, as they were observed in different locations and also have a different diet and therefore it is important for the techniques to be scalable to other species.

Regarding seasonality, some teams noticed that the location of the clusters changed during different seasons, and most sightings happened during summer, which is the breeding season. The change in the clusters across seasons can be seen in Figure 15.

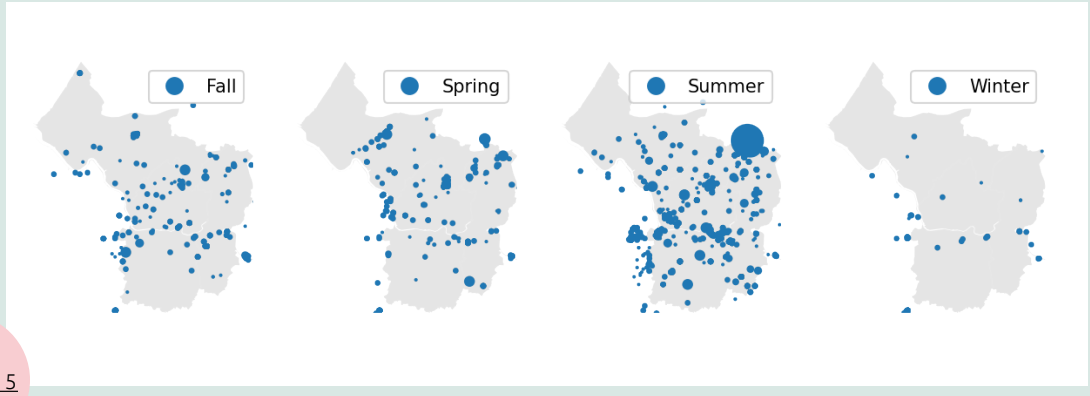


FIGURE 15

Observation of bats in the city of Bristol in different seasons.

PRODUCT

Most of the teams proposed a tool to assist the city officials in finding the location where the lights should be turned off. The proposed main activity was to create suggestions on where to turn off the lights to form a dark ecological corridor through the input of different species and locations. Other features presented by the teams were predicting the change in the population, the output of seasonal corridors, the input of new restrictions, and parameterization of the different constraints and restrictions. The resulting output would be a map of the ecological dark corridors and a list of lights to be turned off and/or dimmed. An example of a dashboard can be found in Figure 16.

One team suggested creating an informative website for the population of the city explaining the importance of the project and a map with the lights projected to be turned off. On this website, the citizens could also provide feedback to the city council and request to turn on the lights in the case of an emergency.

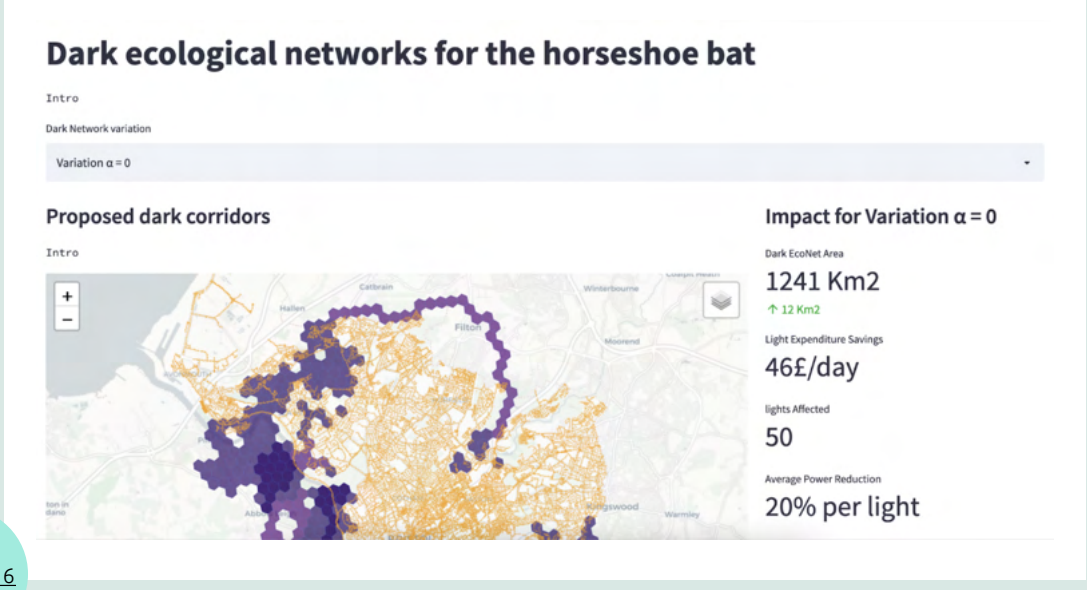


FIGURE 16

Example of a dashboard with proposed dark corridors and the impact on the lights.

SOCIAL IMPACT

The main outcome identified by the teams was to mitigate the impact of ALAN and increase the city's biodiversity without impacting humans. It was proposed to measure the following impact metrics:

- Number of horseshoe bats observed in the city and its surroundings
- Artificial light pollution by measuring the brightness of the sky
- Saved CO2 from shutting down lights
- Financial savings on electricity from shutting down lights

Besides these, safety metrics should be monitored to ensure that there is no negative impact from the implementation of the solution.

The teams estimated that by implementing this proposal the bat activity could be increased by 10%, saving about 185 tons of emitted CO2 and £25,000 in electricity from shutting down lights.

Conclusions

This report is a summary of the work done by more than 200 participants over the course of several months. As an outcome of the competition, 70 technical reports were gathered on 7 different challenges. All the technical reports include code, which we hope can be used to prototype and develop the solutions further in the future. It is essential to mention that these reports differ in quality, which can be assessed by the position of the respective team on the leaderboards.

The challenges provided to the participants were also very varied in terms of field of study - optimization, time series forecasting, computer vision, geospatial analytics, among others - the size of the datasets and the type of available data. On top of it all, the data was provided by real-world institutions or from open data portals, which posed a true test and challenge to all participants in terms of data-centric development.

Looking into the future, WDL is making the reports from each challenge even more accessible to a non-technical audience. We have launched our own Social Impact Hub, which will compile and openly share all the knowledge and code from all WDL editions in an intuitive and easily-navigable way. Besides that, WDL actively supports teams that want to develop their ideas further and ensure that the methodologies presented are scientifically rigorous.

We hope these results can spark future research directions and provide feedback to cities on how data can be leveraged to solve their challenges.

WDL was nominated as a **Global Top 100 AI solution for Sustainable Development** by the International Research Center on Artificial Intelligence under the auspices of UNESCO (IRCAI).

Interested in our work or would like to submit a challenge?

Get in touch with us at

hi@worlddataleague.com

References

STAGE 1 - ENVIRONMENT

Predict Waste Production for its Reduction

[1] World Bank. "A Global Snapshot of Solid Waste Management to 2050". Available at: https://datatopics.worldbank.org/what-a-waste/

[2] Quintili, A., Castellani, B., 2020. The Energy and Carbon Footprint of an Urban Waste Collection Fleet: A Case Study in Central Italy. MDPI. Available at: https://mdpi.com/2313-4321/5/4/25/pdf

[3] Government of Austin, Texas. “Zero Waste by 2040”. Available at: https://www.austintexas.gov/zerowaste

[4] Environment Agency of the UK Government. “Pollution inventory reporting – incineration activities guidance note”. Available at: https://bit.ly/3QEsYUg

Air Quality Prediction in Busy Streets

[5] World Health Organization. “Air pollution”. Available at: https://www.who.int/health-topics/air-pollution#tab=tab_1

[6] Amsterdam Air Quality Institute. “Air quality in Amsterdam”. Available at: https://www.iqair.com/netherlands/north-holland/amsterdam

[7] Our Word in Data. “Deaths from air pollution, 1990 to 2019”. Available at: https://ourworldindata.org/grapher/air-pollution-deaths-country?tab=chart&country=~NLD

[8] Szarata, A., Nosal, K., Duda-Wiertel, U. and Franek, L., 2017. The impact of the car restrictions implemented in the city centre on the public space quality. Transportation Research Procedia, 27, pp.752-759.

References

STAGE 2 - TRANSPORTATION & MOBILITY

Optimization of public transport routes during road interruptions

[9] American Public Transportation Association. "Public Transportation Facts". Available at: https://www.apta.com/news-publications/public-transportation-facts

Predicting the flow of people for public transportation improvements

[10] Nasrudin, Na'asah & Rostam, Katiman & Mohd Noor, Harifah. (2014). Barriers and Motivations for Sustainable Travel Behaviour: Shah Alam residents' Perspectives. Procedia - Social and Behavioral Sciences. 153. 10.1016/j.sbspro.2014.10.084.

[11] TomTom 2022, Tom Tom Traffic Index. Available at: https://www.tomtom.com/en_gb/traffic-index/

[12] Association of Central Oklahoma Governments (2013, November). Why Transit Matters: The Environmental Benefits of Public Transportation. Available at: https://www.acogok.org/why-transit-matters-environment/

Optimization of soft-mobility drop-off points

[13] Şengül, B., & Mostofi, H. (2021). Impacts of E-Micromobility on the Sustainability of Urban Transportation—A Systematic Review. Applied Sciences.

[14] TomTom 2022, Tom Tom Traffic Index. Available at: https://www.tomtom.com/en_gb/traffic-index/

References

SEMI-FINALS - SAFETY

Predicting a safety score for women in Costa Rica

[15] UN Women. "Safe Cities and Safe Public Spaces: Global results report". Available at: https://bit.ly/3QT02rd

[16] Observatory for Gender-based Violence Against Women of the Costa Rica Government. Available at: https://observatoriodegenero.poder-judicial.go.cr/

[17] Pinkerton Consulting & Investigations. “Pinkerton Crime Index Methodology”. Available at: https://pinkerton.com/products/pinkerton-crime-index/methodology

[18] Chalfin, A., Hansen, B., Lerner, J. et al. Reducing Crime Through Environmental Design: Evidence from a Randomized Experiment of Street Lighting in New York City

FINALS - BIODIVERSITY

Identification of Dark Ecological Corridors

[19] Kévin Barré, Arthur Vernet, Clémentine Azam, Isabelle Le Viol, Agathe Dumont, Thomas Deana, Stéphane Vincent, Samuel Challéat, Christian Kerbiriou, Landscape composition drives the impacts of artificial light at night on insectivorous bats, Environmental Pollution, Volume 292, Part B, 2022, 118394, ISSN 0269-7491, https://doi.org/10.1016/j.envpol.2021.118394

[20] Bo Luo, Rong Xu, Yunchun Li, Wenyu Zhou, Weiwei Wang, Huimin Gao, Zhen Wang, Yingchun Deng, Ying Liu, Jiang Feng, Artificial light reduces foraging opportunities in wild least horseshoe bats, Environmental Pollution, Volume 288, 2021, 117765, ISSN 0269-7491, https://doi.org/10.1016/j.envpol.2021.117765