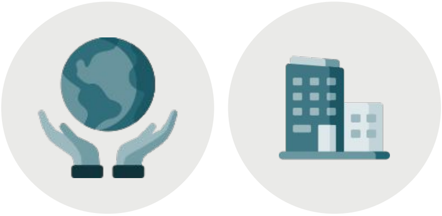
Applying geospatial data for Machine Learning
with a focus on social good

Paulo Maia
Data Scientist at NILG.AI
DSSG Webinar
30th June 2020

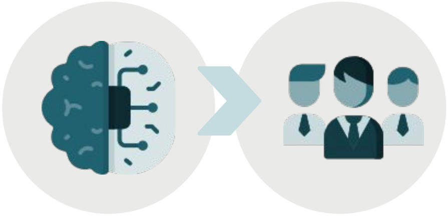
Unlocking business capabilities

using data intelligence

Consulting company
in **Artificial Intelligence**



We help small-to-large companies to make informed decisions promoting
efficient opportunity handling.



We **adapt the technology** to your business, so you can seamlessly **integrate data** in your daily decisions.

Our Network

Clients

Healthcare

Telco

Fintech

Chemistry

Utilities

Marketing

Partners

COLETIV

pwc

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

DEI DEPARTAMENTO DE ENGENHARIA INFORMÁTICA

UPTEC

nilg.ai

Meet the team

Kelwin Fernandes
PhD Computer Science
CEO

Nohelia González
M.Arch
COO

Tiago Freitas
MSc Bioengineering
Data Scientist

Ricardo Azevedo
MSc Computer Science
Data Scientist

Francisca Morgado
MSc Bioengineering
Data Scientist

Paulo Maia
MSc Bioengineering
Data Scientist

nilg.ai

01

Spatiotemporal data types

Spatiotemporal data?

Spatiotemporal (ST) data is defined as data with both a **spatial and temporal component.**

Examples

- Remote Sensing Data
- Route taken by a transport
- Measurements done by a sensor in the street
- Medical imaging (fMRI exam)
- Videos

These are applicable in several domains, such as **environmental and climate, public safety and human mobility.**

Sources: https://arxiv.org/pdf/1906.04928.pdf

Why we should treat ST data differently

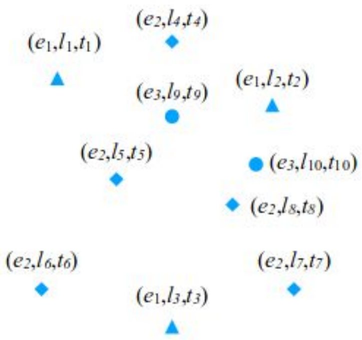
Traditional Machine Learning methods (based on tabular data) **don't work as well in spatiotemporal data.**

- Typically, spatiotemporal (ST) data is in **continuous space**
- Patterns of ST data typically contain **spatial and temporal properties** - hard to capture by traditional methods
- Data samples are **not independently generated**
- Traditional methods are very reliant on **manual feature engineering**, which is hard to do on spatiotemporal data.

nilg.ai

Sources: https://arxiv.org/pdf/1906.04928.pdf

Data Types - Event



Discrete events occurring at **point locations and times**

(crime events, traffic accidents, disease outbreaks, social events and trending topics, civil protection)

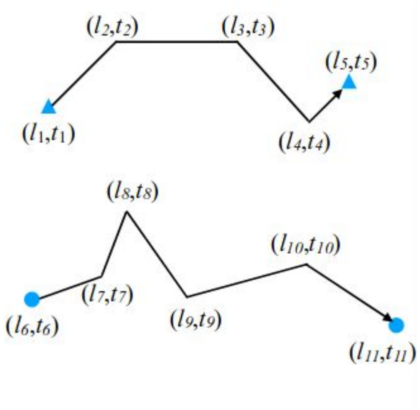
e_i - Event type
l_i - Location
t_i - Time

Events | Operatives | Ground assets | Aerial Assets | Important events | Today's Events | Advanced search

Event Nr.	Status	District	Municipality	Parish	Locality	Day/Time	Nature			
2020130093053		PORTO	PENAFIEL	Termas de São Vicente	TERMAS DE SÃO VICENTE	2020.06.26 14:56	Trauma	2	1	0
2020130093050		PORTO	GONDOMAR	Fânzeres e São Pedro da Cova	FÂNZERES E SÃO PEDRO DA COVA	2020.06.26 14:55	Patrulhamento, Reconhecimento e Vigilância	5	1	0
2020130093041		PORTO	MATOSINHOS	Custóias, Leça do Balio e Guifões	MATOSINHOS	2020.06.26 14:46	Detritos não confinados	12	3	0

Sources: http://www.prociv.pt/en-us/SITUACAOOPERACIONAL/Pages/default.aspx?cID=14, https://arxiv.org/pdf/1906.04928.pdf

Data Types - Trajectory

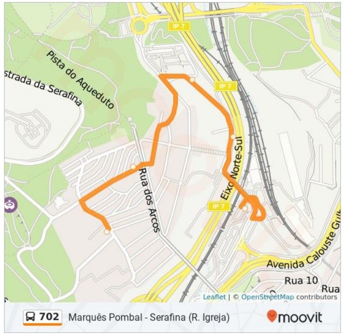


Denote the paths traced by **bodies moving in space over time** (moving route of transports).

Collected by **sensors in a moving object**, for instance.

Represented by a sequence:

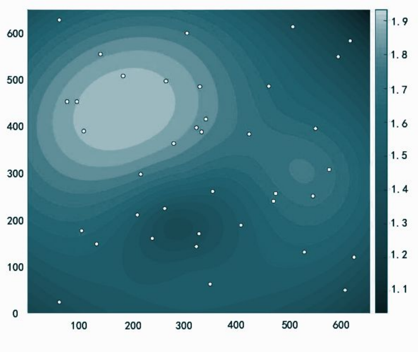
(l_i, t_i) where **l_i** is the location and **t_i** is the time.

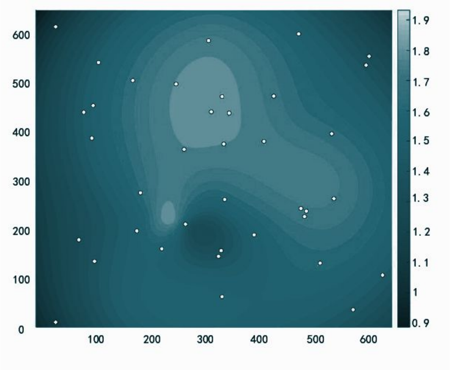


nilg.ai

Sources: https://carris.tecmic.com/index.html, https://arxiv.org/pdf/1906.04928.pdf

Data Types - Point Reference





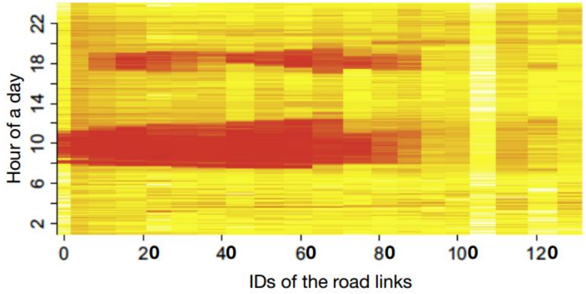
Measurements of a **continuous ST field** (temperature, vegetation, population) over a **set of moving reference points in space and time** - e.g. weather balloons.

Tuple $(\mathbf{r}_i, \mathbf{l}_i, \mathbf{t}_i)$ with measurement value **m**

$\mathbf{r}_i$ - Sensor
$\mathbf{l}_i$ - Location
$\mathbf{t}_i$ - Time

Sources: https://arxiv.org/pdf/1906.04928.pdf

Data Types - Raster Data



ST fields recorded at **fixed locations and time points** (the previous datatype - point reference - had **changing locations**)

e.g. air quality data, traffic flow/car speed

For *m* fixed locations and *n* timestamps, we can present the data as a matrix R (*mxn*).

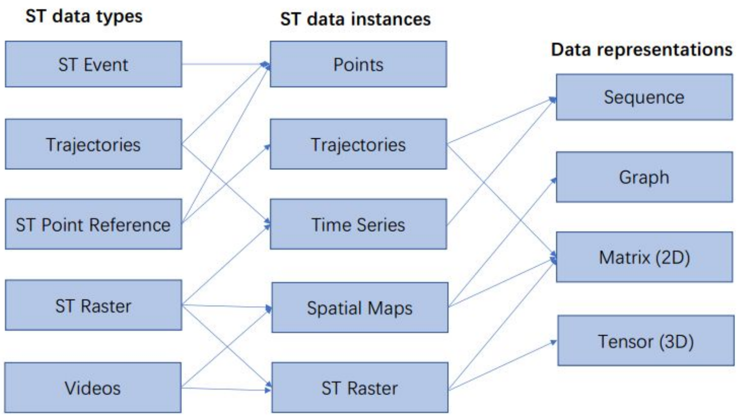
Each entry $\mathbf{r_{ij}}$ is the measurement at location $\mathbf{s_i}$ at timestamp $\mathbf{t_j}$.

Sources: https://qualar1.apambiente.pt/qualar/index.php?page=2, https://arxiv.org/pdf/1906.04928.pdf

02

Treating this data

Representation types



There is not a **best** way of representing ST data.

Traffic Flow Prediction

- Traffic flow graph
- Cell region-level traffic flow matrix

Wind Forecasting

- Represent sensor data as 2D matrices containing wind intensity: (location, timesteps)
- Can also represent as 3D tensor: row region sensor ID, column region sensor ID, timestamp.

Sources: https://arxiv.org/pdf/1906.04928.pdf

Modeling approaches in the literature

	Trajectories	Time Series	Spatial Maps (Image-like data & Graphs)	ST Raster
CNN	[24], [67], [103], [117], [150]		[11], [154], [199], [152], [100], [31], [139], [148], [184], [80], [69], [15], [72], [200], [113], [54], [68]	[188], [12], [123], [141], [106], [74], [131], [149], [116], [128], [128], [76], [78]
GraphCNN			[85], [155], [94], [111], [144], [22], [92], [175], [44], [8], [85], [155]	
RNN(LSTM,GRU)	[42], [77], [165], [99], [91], [163], [35], [159], [64], [38], [135], [181], [88], [81], [190], [37], [169], [166], [41], [65], [192]	[126], [27], [177], [90], [23], [89], [178], [17], [179], [101], [97], [14], [34]	[125], [107], [156], [2], [3], [39], [62], [162]	[23]
ConvLSTM			[1], [98], [161], [198], [151], [73], [201], [70], [147]	
AE/SDAE	[115], [197], [13]	[55], [167], [104]	[32], [16], [191], [48], [52], [182]	
RBM/DBN	[117]	[136]		[140], [58], [66]
Seq2Seq	[82], [170], [20], [171]	[90], [89]		
Hybrid	[164], [142], [108]	[96], [59]	[189], [30], [19], [6], [174], [187], [84], [109], [134], [49], [176]	[105], [127]
Others	[36], [10], [46], [195], [26], [193], [168]	[124], [93]	[133], [145], [202], [21], [183], [146], [79], [43], [185], [186], [132]	[122], [63], [71]

Recurrent Networks
- Mostly used for trajectories and time series

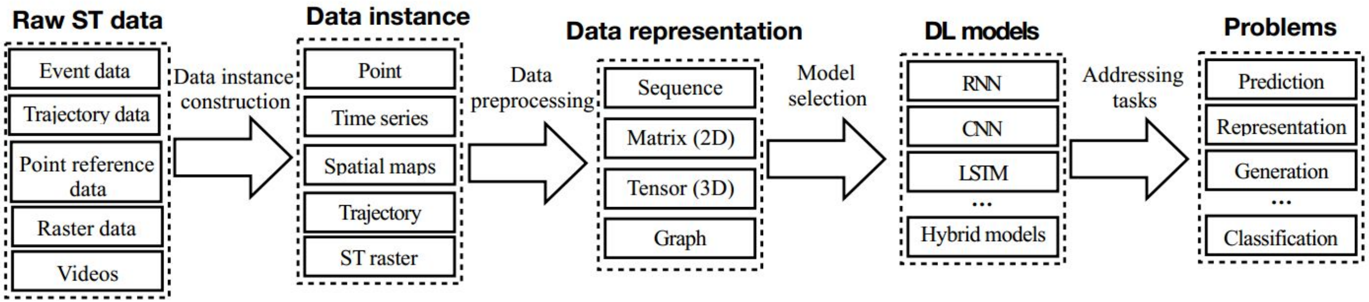
CNNs
- Mostly used for map-like data (spatial maps/ST Raster)

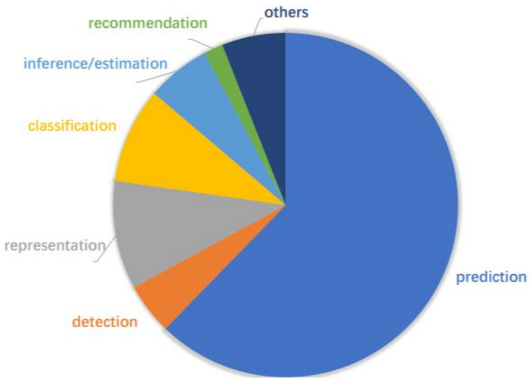
GraphCNN
- Used for graph data

nilg.ai

Sources: https://arxiv.org/pdf/1906.04928.pdf

Pipeline for (most) ST data problems





Prediction: predict the future observations of the ST data based on its historical data.

Representation: learn the abstract and useful representations of the input data to facilitate downstream data mining or machine learning tasks.

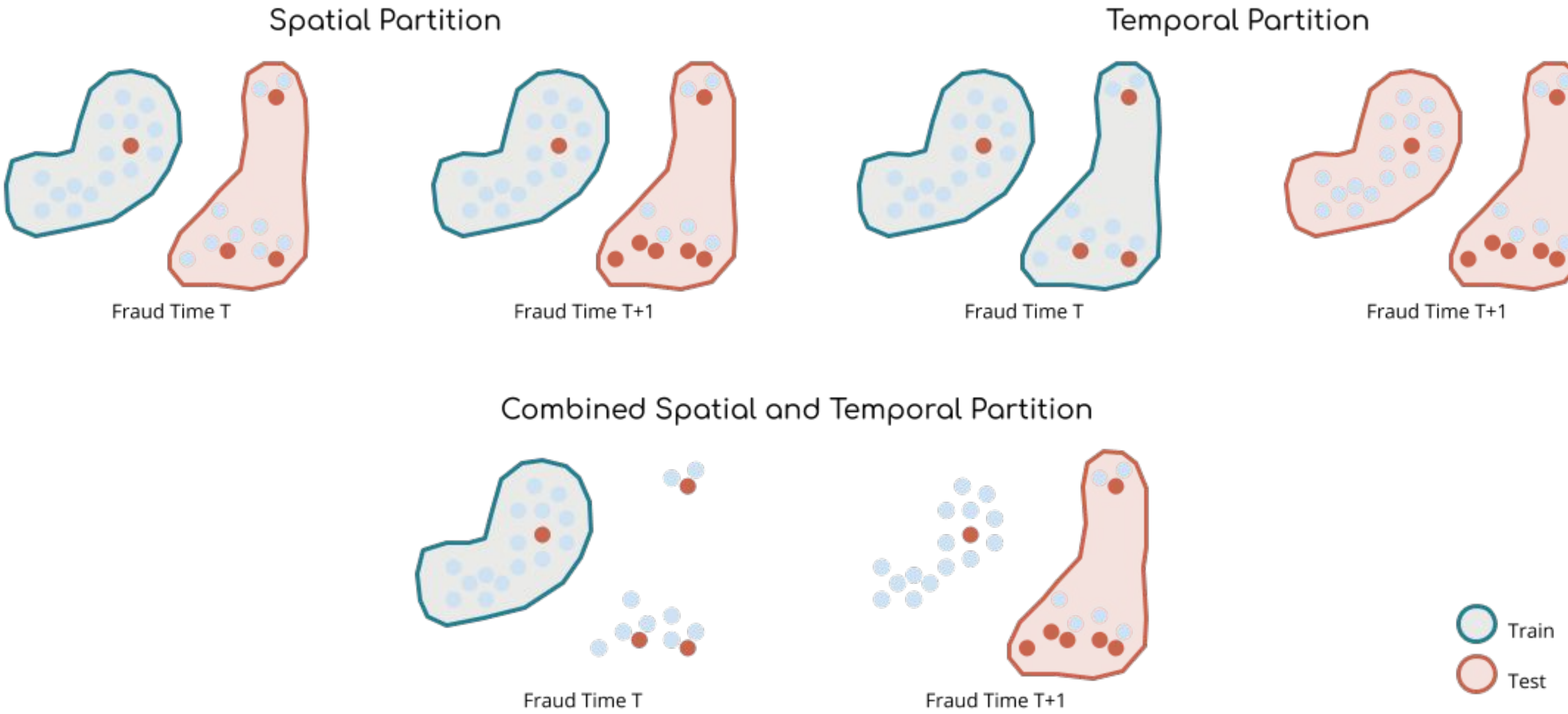
Classification: separate geospatial data in classes.

Estimation: information inference.

Sources: https://arxiv.org/pdf/1906.04928.pdf

nilg.ai

Model evaluation for geospatial problems

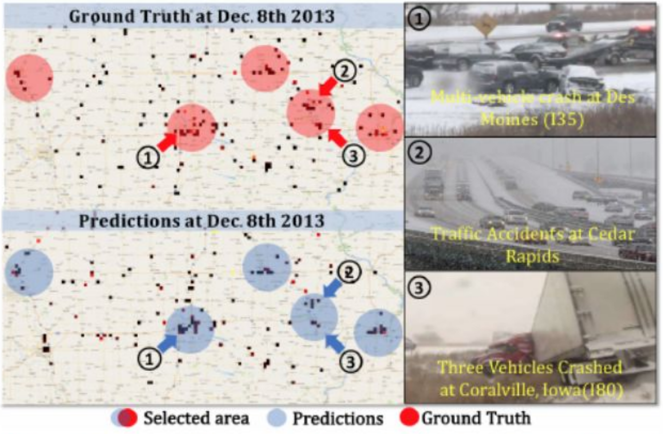


Sources: https://nilg.ai/blog/202006/who-is-tampering-your-meters-fraud-detection-in-utilities/

02

Use Cases

Traffic Accident Count Prediction (i)

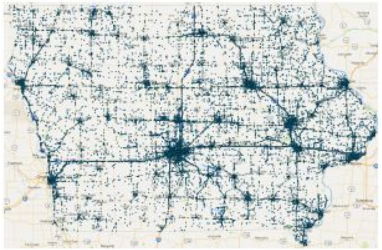


- Has the goal of improving transportation, public safety and safe routing.
- Challenging:
  - **Rareness** of accidents in space and time
  - **Spatial heterogeneity** (rural vs urban environment have different accident factors).
- Commonly formulated as a **classification** or **regression** problem
  - Will there be an accident in (e_i, t_i)?
  - Accident count in (e_i, t_i)?

nilg.ai

Sources: https://dl.acm.org/doi/pdf/10.1145/3219819.3219922

Traffic Accident Count Prediction (ii)



(a) Visualization of Traffic Accidents

(b) Rainfall Map

(c) RWIS Observation Stations

Data Sources

- **Target:** Crash Data
- **Auxiliary:** Road Networks
- **Features:**
  - Satellite Images
  - Traffic Camera Data (i.e. number of vehicles)
  - Rainfall Data
  - RWIS Stations (Temperature and Wind)

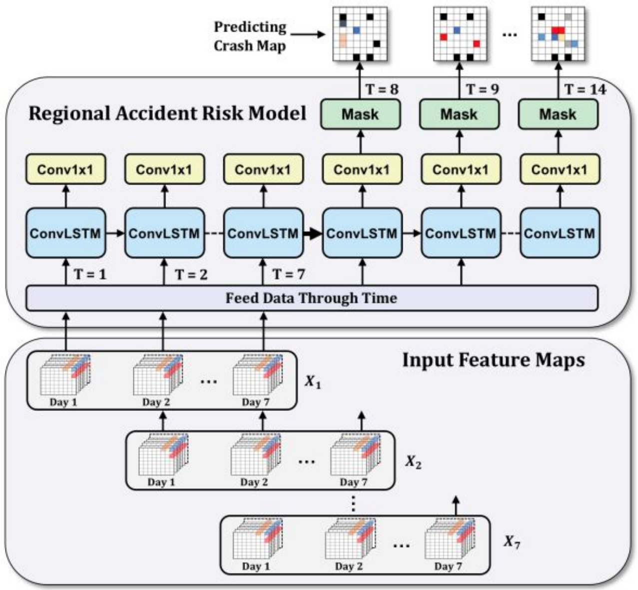
Formulation

- **Data Instances**
  - Map is partitioned in a square spatial grid S (elements $\mathbf{s}_i$).
  - Data Sources above are saved as 3D tensor ($\mathbf{s}_i$, $\mathbf{t}_i$)
  - Map is masked according to road networks
- **Target**
  - Predict the total number of accidents in a given time window, for cell $\mathbf{s}_i$.

nilg.ai

Sources: https://dl.acm.org/doi/pdf/10.1145/3219819.3219922

Traffic Accident Count Prediction (iii)



Regional ConvLSTM Model

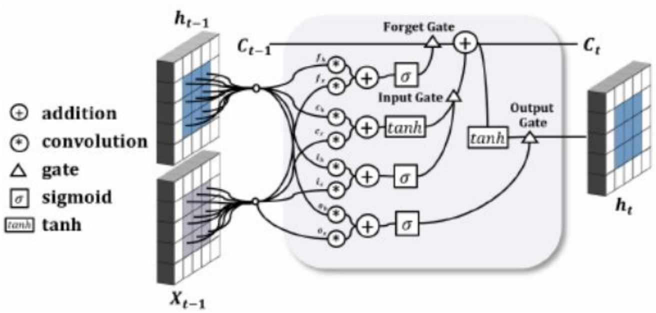


Figure 3: The inner structure of a ConvLSTM cell.

Training and Testing

The last 7 days are predicted based on the data in the first 7 days.

Amount of days chosen is related with the human activity **weekly pattern.**

The first 7 years are used for **training** the model, and the last **year** for testing (partition by time)

Sources: https://dl.acm.org/doi/pdf/10.1145/3219819.3219922

Traffic Accident Count Prediction (iv)

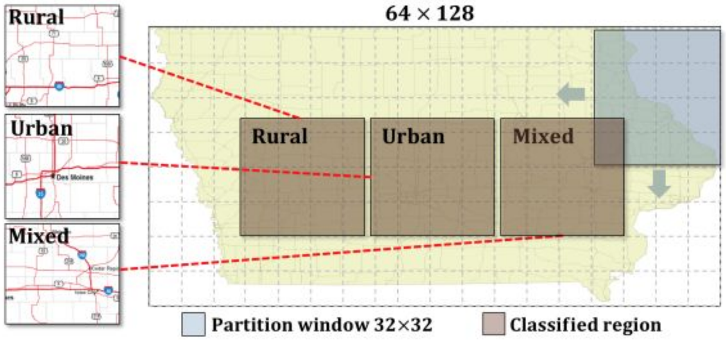


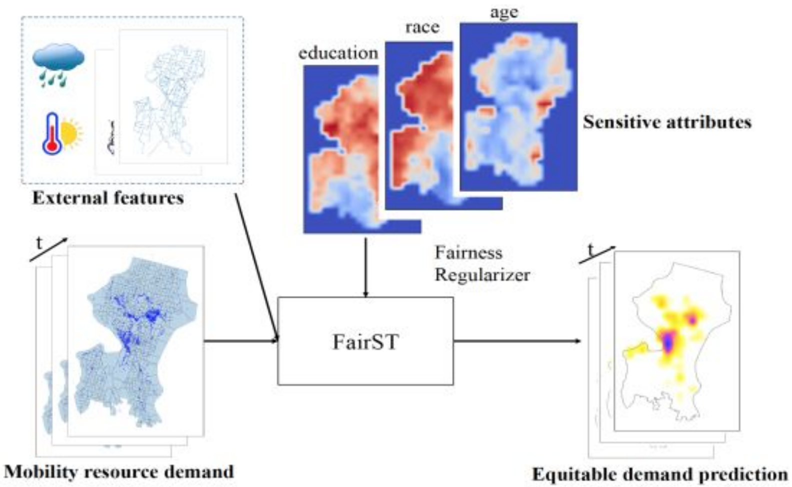
Figure 5: Map partitioning of spatial ensemble model (stride=16).

For dealing with spatial heterogeneity, the authors **learn a model for each window**, and ensemble all the predictions.

Sources: https://dl.acm.org/doi/pdf/10.1145/3219819.3219922

Fair Demand Prediction in Mobility Systems (i)

- New mobility systems (e.g. bike-sharing) offer affordable transport options for citizens.
- Current demand prediction models do not deal with social disparities: less demand **might not mean less interest**, but an area with **disadvantaged groups**!



Sources: https://dl.acm.org/doi/pdf/10.1145/3347146.3359380

Fair Demand Prediction in Mobility Systems (ii)

- How do we incorporate fairness in this approach?

Individuals of different groups must have access to the same resources.

Group Fairness: The disadvantaged group must experience similar predicted outcomes as the advantaged group

Vertical Equity: Transportation policies must favor the disadvantaged groups

Two fairness metrics added, measuring the **gap** between **mean demand per capita** across two groups, over a certain period of time.

Region-based Fairness Gap: each geographic region has a categorical label (e.g. Caucasian)

Individual-based Fairness Gap: the group label is numeric (e.g. % of Caucasian)

Sources: https://dl.acm.org/doi/pdf/10.1145/3347146.3359380

Fair Demand Prediction in Mobility Systems (iii)

- Loss Function is Modified to add the following terms:

Regression Loss, for penalizing wrong predictions in terms of **demand value** (Mean Absolute Error)

Region Fairness Loss, for penalizing the model when the demand per capita in a region containing both groups **is not the same**

Individual Fairness Loss, for penalizing the model when the demand per capita in a region containing both groups **does not favor the disadvantaged group** (weighted).

nilg.ai

Sources: https://dl.acm.org/doi/pdf/10.1145/3347146.3359380

City Inflow/Outflow Prediction (i)

- Crowd flow prediction is important for managing **traffic and public safety.**
  - Measured by the number of pedestrians + cars + people traveling on public transportation systems.
  - This paper uses **GPS data** for measuring the number of pedestrians.

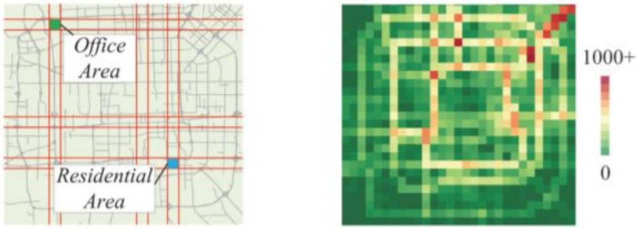
- **Challenging** problem due to:
  - **Spatial dependencies** (outflow in a region affects inflow in another)
  - **Temporal dependencies** (traffic congestion at 8 am affects typical traffic at 9 am, and people's routine changes throughout the year)
  - **External factors** (e.g. weather)

(a) Inflow and outflow (b) Measurement of flows

Source: https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/viewPaper/14501

City Inflow/Outflow Prediction (ii)

- Problem and label definition



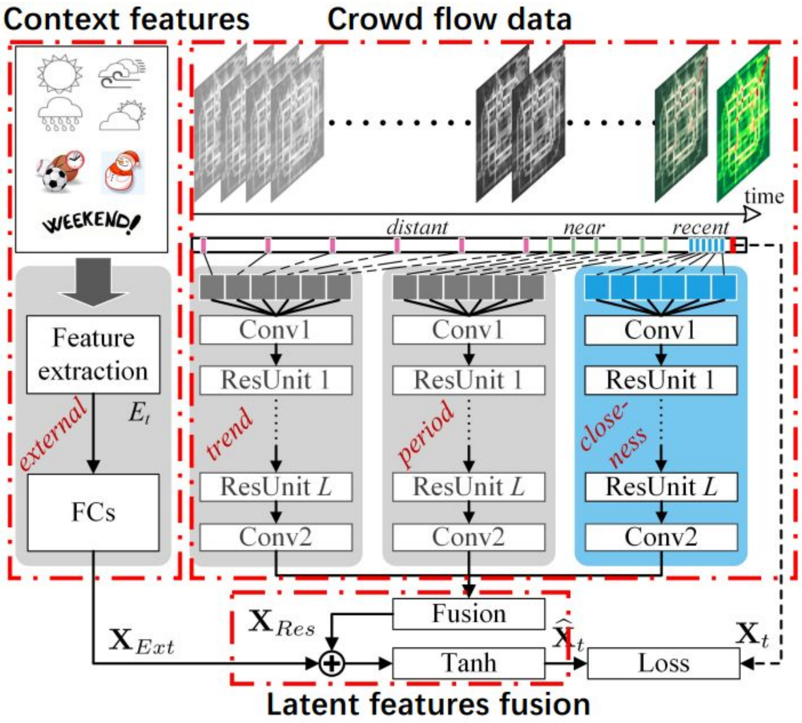
(a) Grid-based map segmentation (b) Inflow matrix

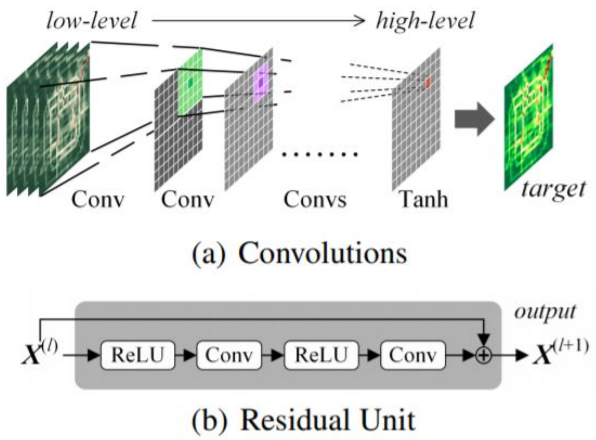
The authors define an **inflow** and **outflow** 2D matrix, for each **timestep**, after dividing the map in a grid.

Given the **historical observations** up to point **T-1**, they want to predict the count of inflow and outflow at point **T**.

Source: https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/viewPaper/14501

City Inflow/Outflow Prediction (iii)





Spatiotemporal ResNet

nilg.ai

Source: https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/viewPaper/14501

Concluding…

There are **many ways** spatiotemporal data can be used for AI applications.

There is not much (if any?) research done using open data in Portugal. **So why not start with it?**

http://centraldedados.pt/

DADOS ABERTOS
em Portugal

http://dadosabertos.pt/

I'm sure **DSSG** would give you some exposure and help you find **motivated volunteers**.

(well, I'm biased)

Applying geospatial data for Machine Learning

with a focus on social good

Paulo Maia
Data Scientist at NILG.AI
DSSG Webinar
30th June 2020