A Global Classification Model of Cities Using ML

Doron Hazan · Mohamed Elhabashy · Mohanned ElKholy · Omer Mousa · Norhan Bayomi · Matias Williams · John Fernandez
Massachusetts Institute of Technology

Motivation

Cities are both the drivers of climate change and the major component of the solution

- Yet, many cities are lacking direction toward a climate-positive and sustainable future
- Many cities are relatively small and lack the resources to know what solution set is most appropriate for them
- Many cities do not know how to connect to other cities to share their stories and journey toward sustainability

Goals & Challenges

Goals

1. Represent a comprehensive set of cities around the world: 9000 cities worldwide
2. Design a novel way to classify cities and cluster similar ones
3. Derive a starting point for pathways for sustainable urban growth, resource efficiency, and climate change solutions

Challenges

1. Severe lack of data (for 9000 cities)
2. So far, to our understanding, no known clustering mechanism for cities on this scale

Our Approach

Generate novel metabolic dataset for 9000 cities & design a novel way to cluster them

Generate Novel Metabolic Dataset

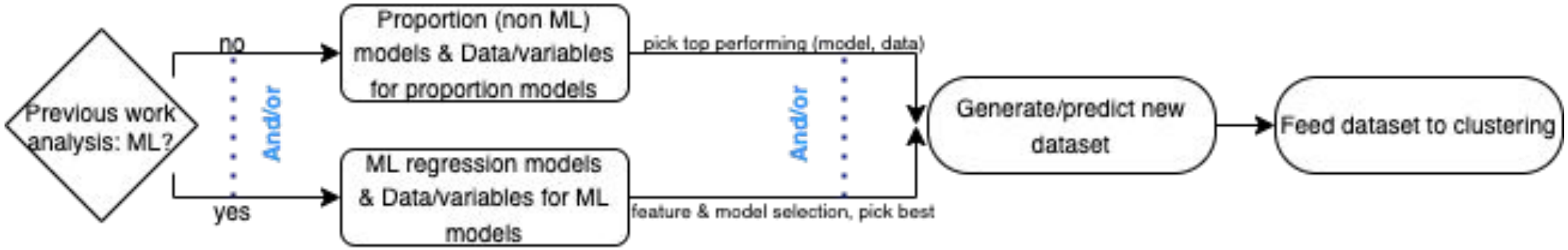
1. Define *resource use* as a characteristic of cities, and predict resource use values for 9000 cities through:
    - Energy consumption
    - Water consumption
    - Food consumption
2. Benchmark results
3. Feed predicted data to clustering algorithm

Cluster Novel Metabolic Dataset

1. Design a three-fold clustering algorithm
2. Benchmark novel algorithm vs baseline clustering approaches
3. Analyze & interpret clustering results

Decision Flow Process

From generating data to clustering



Data

Energy Consumption (average electricity use per capita in kWh)

- N = 155 cities
- Satellite data (light radiance) and population data

Water Consumption (average liters per capita per day)

- N = 172 cities
- Precipitation, population, water price (retail), land area, average temperature (2m above ground)

Food Consumption (average daily consumption of calories per capita)

- N = 1220 cities
- Population, national average daily consumption of calories

Energy Model

We found that city light radiance captured in satellite images approximates well the energy consumption of cities, thus, we employed a city light radiance - population proportion model to capture energy consumption globally

$$E_i^{Capita/City} = \frac{L_i^{City}}{L_j^{Country}} * \frac{P_j^{Country}}{P_i^{City}} * \overline{E}_j^{Capita/Country}$$

Where:

L_i corresponds to light radiance, P_i to population, E_i to energy consumption per city i, and $\overline{E_j}$ to National Average Energy consumption for country j that includes city i.

Water Model

The model that was picked to estimate water consumption was Extremely Randomized Trees (ERT), which is similar to other tree based ensemble algorithms such as random forests, but was found to perform and generalize better.

Unlike random forests, ERT's use the same training set for training all trees and split a node based on both variable index and variable splitting value, while random forests only splits by variable value.

This makes ERTs both more computationally efficient and generalizable than random forests - which is crucial in our setting since we predict 9,000 values using solely 172 data points.

Water Model – Model Selection

Model Benchmark Water			
Model/Metric	MAPE	R^2	Ratio Score
Linear Regression	26.5%	0.257	38%
Ridge regression	26.5%	0.255	38%
K Nearest Neighbors	20.3%	0.409	27.8%
Support Vector Machines (linear)	25.8%	0.259	43.6%
Support Vector Machines (polynomial)	58%	0.26	35.4%
Support Vector Machines (radial)	19.4%	0.497	25.7%
Decision Trees	28.4%	0.133	25.7%
Random Forest	15.3%	0.589	**19.7**
Extremely Randomized Trees	**13.6%**	**0.625**	20.3%
Extreme Gradient Boosting	14.8%	0.542	21.9%
Multi-layered Perceptron	30.8%	0.274	37.1%

Food Model

Findings from the literature indicated that the city's food consumption is highly correlated with population. The population proportion output is then used to estimate food consumption via linear regression:

$$F_i^{Capita/City} = \widehat{\beta_0} + \widehat{\beta_1} * \left(\frac{P_j^{City}}{P_i^{Country}} * F_j^{Country} \right) + \varepsilon_i$$

Where:

P_i and F_i corresponds to population and food consumption for city i, respectively. F_j for food consumption for country j that includes city i. $\beta's$ are regression coefficients and ε_i is the error rate for city i.

Benchmarking

The predictions of energy, water, and food consumption of 9000 cities enabled constructing a novel data set that entails resource consumption of cities on a global scale.

Since this data is novel, we wanted to benchmark it through four different approaches:

1. Standard regression metrics (MAPE and R^2) via k-cross validation (k=10)
2. Statistical characteristics (comparing mean, median, range, variance between the original data and the predicted data)
3. *Ratio Score* metric, which we have developed specifically for this task
4. Qualitative metric: (on going work)

Ratio Score Algorithm

Algorithm 1: Ratio Score

$\mathcal{T} = \{y_1, ...y_N\}$: true testing data
$\mathcal{P} = \{y_1, ...y_N\}$: predictions on testing data
$\mathcal{R} = 0$: Ratio Score
Initialize $\mathcal{R}^i_{true}$: a vector of length $N - 1$
Initialize $\mathcal{R}^i_{prediction}$: a vector of length $N - 1$
for *city name* $i \in \mathcal{T}$ **do**
    for *city name* $j \in \mathcal{T}, j \neq i$ **do**
        Compute $\mathcal{R}^i_{true} = (\frac{\mathcal{T}_i}{\mathcal{T}_j})\ \forall j \in \mathcal{T}, j \neq i$
        Compute $\mathcal{R}^i_{prediction} = (\frac{\mathcal{P}_i}{\mathcal{P}_j})\ \forall j \in \mathcal{P}, j \neq i$
    end
    Compute $\mathcal{R}^i_{error} = \frac{|\mathcal{R}^i_{true} - \mathcal{R}^i_{predicted}|}{\mathcal{R}^i_{true}}$
    Update $\mathcal{R} = \mathcal{R} + \mathcal{R}^i_{error}$
end
Return $\mathcal{R}$

Ratio Score Algorithm – Example (Water consumption)

Data Set Benchmark					
City/Value	True Value	Predicted Value	MAPE	True Ratio	Predicted Ratio
Geneva (Switzerland) Tokyo (Japan)	171 220	188.92 244.93	10.48% 11.33%	**0.78**	**0.77**

Results

Statistical Characteristics

Data Set Benchmark		
Measurement/Dataset	Original Data	Our Novel Predicted Data
Mean (Energy)	3014.533	3075.545
Mean (Water)	165.36	166.08
Mean (Food)	10406618	11267953
Median (Energy)	1582.5	2328.9
Median (Water)	148	160.91
Median (Food)	9011173	9954047
Min (Energy)	158	197.5
Min (Water)	71	96.47
Min (Food)	2758824	4198104
Max (Energy)	26790	23657.35
Max (Water)	538.0	326.57
Max (Food)	34192000	29303875
sd (Energy)	3642.505	3294.618
sd (Water)	71.27	31.18
sd (Food)	5621284	4723401

Regression Metrics + Ratio Score

Dataset/Predictions Benchmark			
Resource/Metric	MAPE	R^2	Ratio Score
Energy	67.7%	0.77	89.5%
Water	13%	0.63	20.3%
Food	22.5%	0.71	30.2%

Clustering

The second step is to use outputs from the resource use prediction models to develop the global classification

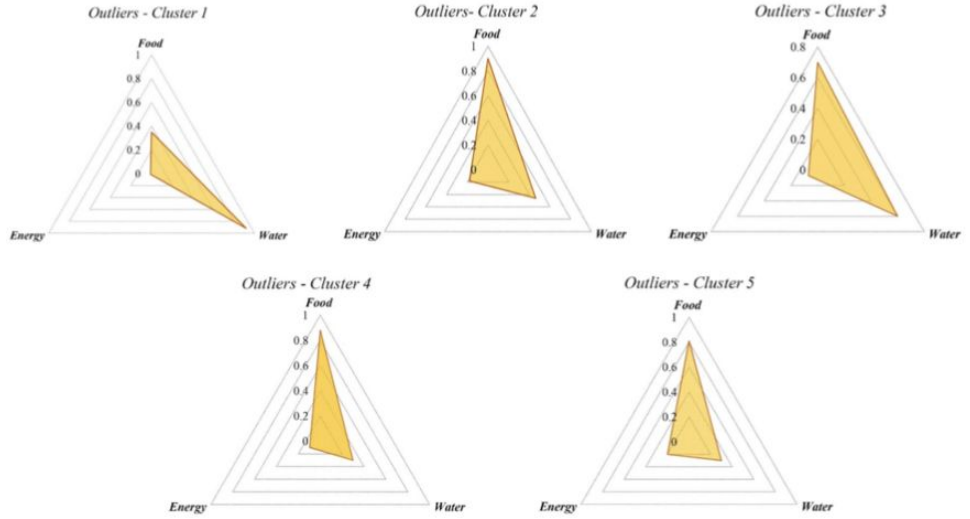
Our proposed clustering pipeline has three main components:

1. Outlier Detection (BACON algorithm): some variables like population and area obey Zipf's Law, where there is an exponential increase in the values of the variables for major cities per country.

2. Training a variational autoencoder (VAE) as a pre-processing step: makes it possible to perform a non-linear transformation that encodes the data into a finer, more separated, and denoised representation.

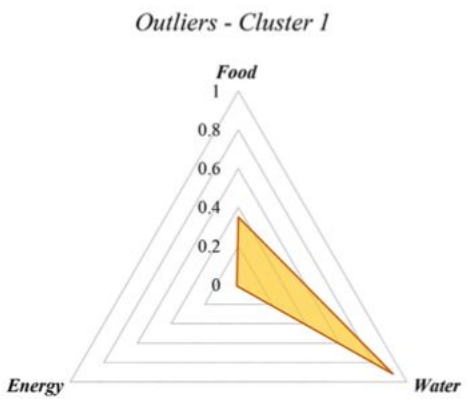
3. Cluster using traditional clustering method (agglomerative clustering): ultimately clusters the 9000 cities based on novel resource-use data

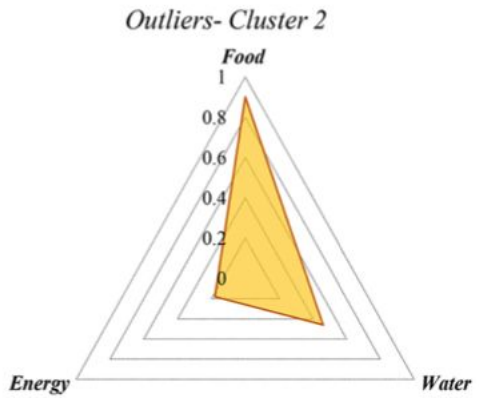
Clustering Results

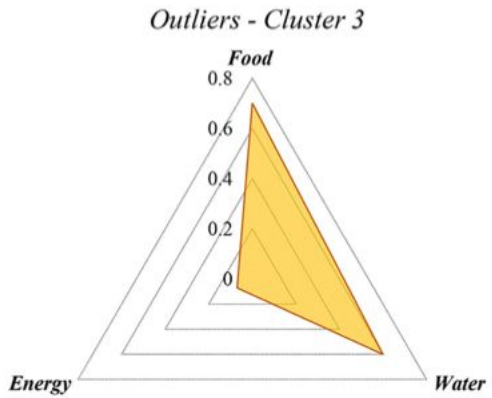


Clustering Algorithm Performance			
Metric/Algorithm	AC only	OD+AC	OD+VAE+AC
Calinski-Harabasz Index	4345.80	2546.07	**42495.66**
Silhouette Coefficient	0.45	0.17	**0.47**

Clustering - Interpretation







For instance, cluster two includes cities at the top 500 range in food consumption and medium range in water use, like Oklahoma (US), Nashville (US), Columbus (US), Buenos Aires (ARG), and London (UK).

Many cities of this cluster are already working together to address climate change under the C40. Each city has drafted plans according to their needs; however, **why couldn't they write plans together when they have similar needs?**

Future Work

1. Other typologies
2. Expand benchmarking
3. Data acquisition
4. Predictions to prescriptions

Thank you!

Please reach out if you are interested in our work!

Contact:
doronh@mit.edu
habashy@mit.edu
mohanned@mit.edu
omermousa@aucegypt.edu

Contact:
nourhan@mit.edu
mwill88@mit.edu
fernande@mit.edu

Groups:

Urban Metabolism Group
https://umg.mit.edu/

Environmental Solutions Initiative
https://environmentalsolutions.mit.edu/