A Global Classification Model of Cities Using ML

Doron Hazan[1], Mohamed Habashy[1], Mohanned ElKholy[1] , Omer Mousa[2], Norhan Bayomi[1], Matias Williams[1] & John Fernandez[1]

[1] *Massachusetts Institute of Technology*, [2] *American University in Cairo*

Abstract

In this work we develop a novel dataset for three key resources use, namely; food, water, and energy, for 9000 cities globally. The dataset is then utilized to develop a clustering approach as a starting point towards a global classification model. This novel clustering approach aims to contribute to developing an inclusive view of resource efficiency for all urban centers globally. The proposed clustering algorithm is comprised of three steps: first, outlier detection to address specific city characteristics, then a Variational Autoencoder (VAE), and finally, Agglomerative Clustering (AC) to improve the classification results. Our results show that this approach is more robust and yields better results in creating delimited clusters with high Calinski-Harabasz Index scores and Silhouette Coefficient than other baseline clustering methods.

Keywords: variational autoencoders, clustering, extremely randomized trees, benchmarking.

Introduction

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. The main limitation to achieving efficient city networks is that many cities are relatively small and lack the resources to know what solution set is most appropriate for them and how to connect to other cities to share their stories and journey toward sustainability. This work addresses this challenge by:

- Using ML to generate a novel resource use dataset for 9000 cities globally
- Designing a novel clustering algorithm to cluster these 9000 cities using our novel dataset

Materials & Methods

We developed a machine learning approach to predict energy, food, and water consumption for a total of 9,000 cities around the world via three distinct pairs of model & data:

Energy Consumption:

$$E_i^{cap/city} = \frac{L_i^{city}}{L_j^{ctry}} \times \frac{P_j^{ctry}}{P_i^{city}} * \overline{E}_j^{cap/ctry} \tag{1}$$

Food Consumption:

$$F_i^{cap/city} = \hat{\beta}_0 + \hat{\beta}_1\left(\frac{P_i^{city}}{P_j^{ctry}} \times F_j^{ctry}\right) + \epsilon_i \tag{2}$$

Water Consumption:

$$W = ERT(precip, temp, price, area, pop) \tag{3}$$

ERT refers to Extremely Randomized Trees model.

Results 1

We utilized a comprehensive set of existing and novel **benchmarking** criteria for our models & data:

- Regression metrics (MAPE, R^2, MAE, etc.)
- Comparing statistical characteristics of predictions and ground truth data
- Ratio Score, a novel metric

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%

Table 2: Data & Model Benchmark

Results 2

Table *1* demonstrates the results of the approaches, the VAE + OD + AC (extracting outliers,106 passing them to VAE and clustering them) produced the highest score for CHI (9.7 times more than just using AC). Thus, our analysis suggests that our novel, three fold clustering method performs better classification on our novel dataset than baseline clustering.

Metric/Alg	AC	OD+AC	OD+VAE+AC
CHI	4345.80	2546.07	**42495.66**
SC	0.45	0.17	**0.47**

Table 1: Clustering Algorithm Performance

Figure 1 shows three example clusters in the outlier group. 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?* This is one way our proposed global clustering approach can help cities address their climate challenges.

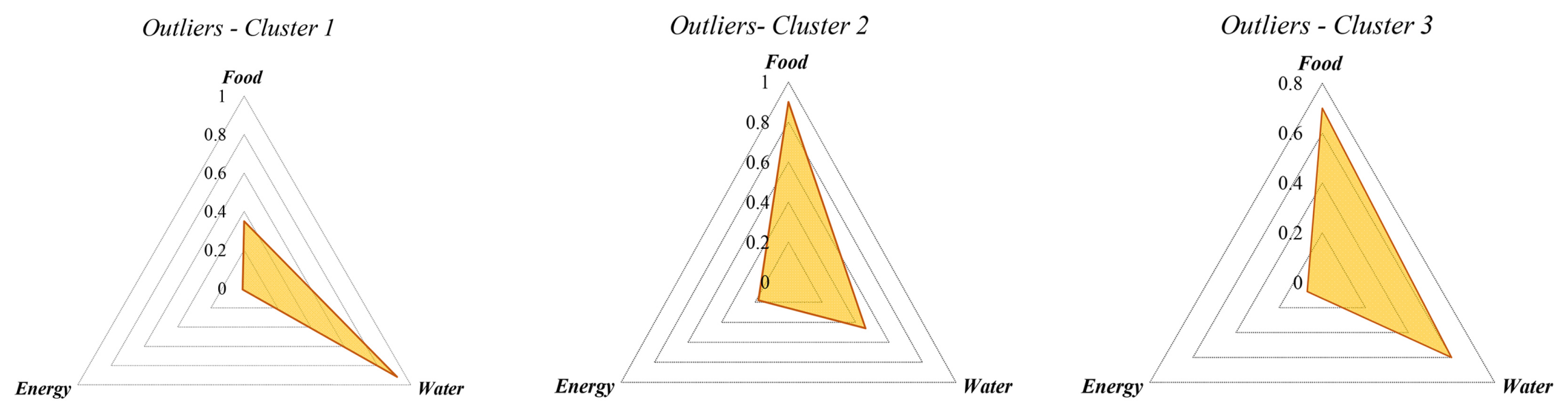

Figure 1: Spider plots for the three example clusters in the outliers group

Conclusion

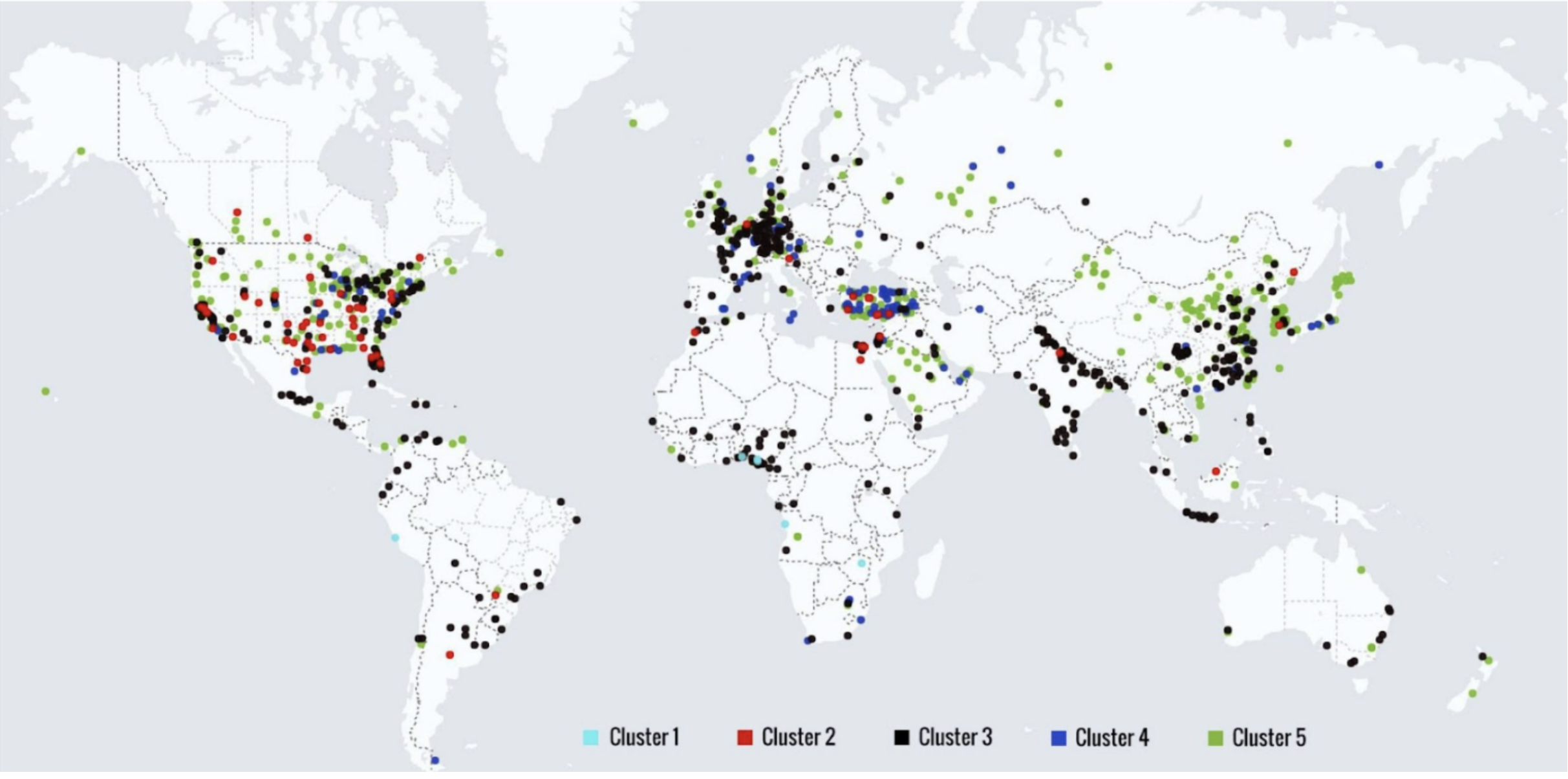

Figure 2: Spatial distribution of the outliers group

The proposed global clustering approach for cities using ML techniques has two practical implications:

- it will provide space for a comprehensive assessment of cities globally and help identify the aggregate contribution of urban areas to global climate challenges.
- global clustering of cities will allow the comparison between cities with similar features and derive pathways for sustainable urban growth, resource efficiency, and climate change challenges.

The data presented in this paper are novel and unique as they are **fulfilling gaps in data scarcity for the majority of cities globally** that limits the opportunities for resource efficiency and sustainable urban growth.

Future Research

Future work includes investigating and identifying a pathway towards sustainability using a global model of cities. In addition, in future work we aim to expand our benchmarking criteria to rigorously evaluate our novel dataset for future ML research use. Moreover, future work should include the analysis of further domains, such as socio-economic status of cities, climate stressors, etc. Data acquisition is needed, and an option of using synthetic data should be explored. Ultimately, we hope that these predictions will be used as prescriptions - decisions - for policy makers towards tackling climate change.

Contact Information

Web https://environmentalsolutions.mit.edu/

Email doronh@mit.edu; nourhan@mit.edu

Phone +1 (857) 389 8891