A Global Classification Model for Cities using ML

Doron Hazan
MIT
doronh@mit.edu

Mohamed Habashy
MIT
Habashy@mit.edu

Mohanned ElKholy
MIT
Mohanned@mit.edu

Omer Mousa
American University in Cairo
omermousa@aucegypt.edu

Norhan Bayomi
MIT
nourhan@mit.edu

Matias Williams
MIT
Mwill88@mit.edu

John Fernandez
MIT
fernande@mit.edu

Abstract

This paper develops a novel data set for three key resources use; namely, food, water, and energy, for 9000 cities globally. The data set 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.

1 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 latest IPCC report has underlined the role of international climate networks between urban centers [1], [2] such as city networks. 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 is problematic as smaller cities face different barriers from their global counterparts [3], and these cities are likely to define urbanization's future, especially in the Global South [4]. This paper directly addresses this dichotomy by using machine learning to develop a global classification approach for cities into various profiles based on quantitative characteristics that can enhance the understanding of urbanization pathways.

Related Work: Most cities' classification studies have mostly focused on hierarchical data-driven methods and do not move beyond comparing a limited number of cities with territorial similarities or development levels [5]. To our knowledge, a clustering approach for cities worldwide has not been fully explored in the literature to date. Such global classification has been challenging due to the lack of available data on a global scale. Thus, this paper develops a novel data set and a broad definition of city boundaries to develop a clustering approach for 9000 cities worldwide. The work presented in this paper is divided into two key components. First, the development of global prediction models for resource use (energy, water, food) for 9000 cities globally to fulfill the current gap in data needs to achieve inclusive clustering of cities. Second, a novel clustering approach that performs better than

Submitted to Tackling Climate Change with Machine Learning: workshop at NeurIPS 2022.

baseline clustering and identifies possible city networks that can aid in the global climate change discussion needs and resource efficiency opportunities.

2 Data and Methods

2.0.1 Data

In this paper, we developed a machine learning approach to predict energy, water, and food consumption for a total of 9,000 cities around the world. This data - for a comprehensive 9000 cities around the world - simply does not exist. Thus, we collected resource data for a subset of cities around the world and processed each dataset to construct our predictive models. For further information regarding the data, methods, and evaluations, refer to appendices A, B. Figure 2 portrays our decision flow process with picking the best resource estimation model.

2.0.2 Methods

Energy Consumption Model: To generate a comprehensive energy estimation model, we used city light radiance as a proxy. We found that city light radiance captured in satellite images approximates well the energy consumption of cities, so we employed a city light radiance proportion model that estimates the consumption of a specific city based on the city's light radiance percentage of the whole country multiplied by the country's population proportion and total energy consumption following previous work in [6], [7], [8] using equation 1. Our model selection analysis is presented in appendix A.2

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

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

Water Consumption Model: To remedy the lack of observational data for water consumption, we limited our model search space to models that work well with little data. Inspired by Fan et al. [9], we performed feature selection to pick a subset of features that first, accurately depict water consumption, second, are generalizable enough to be used for 9,000 cities. For example, the number of washing machines per household was used in [9], but could not be used for 9,000 cities simply because it is unavailable for this large set of cities. The model that was picked to estimate water consumption was Extremely Randomized Trees (ERT) [10], which is similar to other tree based ensemble algorithms such as random forests, but was found to perform and generalize better. The ERT model incorporated total population, land area (from [11]), precipitation (from [12]), temperature (from [13]), and water price (from [14]) which corresponded to our two aforementioned conditions. For further information refer to appendix A.3.

Food Consumption Model: Findings from the literature indicated that the city's food consumption is highly correlated with population. These findings prompted us to develop a population proportion model to estimate food consumption following the same approach in [8]. We define food consumption as the average daily consumption of calories. The population proportion output is then used to estimate food consumption via linear regression as seen in equation 2. For further information, refer to appendix A.4.

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

P_i and F_i correspond to population and food consumption, respectively. β's are regression coefficients and ϵ_i is the error for city i. F_j and P_j correspond to food consumption and population for country j that includes city i

2.0.3 Initial Results and Evaluation

The predictions of our models correspond to an estimation of energy, water, and food consumption across 9,000 cities around the world. These predictions form a novel dataset that leverages the power of machine learning research. To quantify the performance of each model (i.e. our new dataset), we used three benchmarks: standard regression metrics (MAPE and R^2 on the out of sample set), statistical characteristics (e.g. comparing mean, median, etc. between the original data and the predicted data), and a Ratio Score metric B.3 which we have developed specifically for this task. Table 1 presents the results for MAPE, R^2, and Ratio Score benchmarks used to compare the ground truth dataset and the predicted set, evaluated on the test set. Our comprehensive benchmark on the new dataset and models' performance is described in appendix B. We discuss he limitations of our predictions in B.4.

Table 1: Evaluation of the predicted dataset on the out of sample test set

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%

2.0.4 Clustering

The second step is to use outputs from the resource use prediction models to develop the global classification. Here, we developed the clustering approach over three key components: 1) outlier detection (OD), 2) encoding with Variational AutoEncoders (VAE) [15], and 3) agglomerative clustering [16]. The intuition behind using an outlier detector is that some variables like population and area obey Zipf's Law [17], where there is an exponential increase in the values of the variables for major cities per country. Therefore, through the outlier detector, we exclude the highest $x\%$ (x is determined empirically) of the data in the attributes of interest. This step has allowed us to separate the data into an outlier and a non-outlier group based on the attributes we want. Next, we apply a VAE based transformation [15] before applying the clustering algorithm. Training a 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. We train the VAE until convergence to reconstruct the input data using KL-Divergence and Mean Square error objectives, then use the encoder part to transform the data (to the VAE latent space). We use the Adam optimizer [18] with a learning rate of 0.001 for training. For the clustering algorithm, we try the standard AC and other standard clustering methods, which are included in appendix C.

3 Initial Clustering Results

To assess the performance of the proposed approach, clusters are evaluated using the Calinski-Harabasz Index (CHI) [19] and the Silhouette Coefficient (SC) [20]. These are standard ways of evaluating clustering as they measure how dispersed/close the points in the clusters are to each other. CHI is unbounded while SC ranges from -1 to 1. Scores for our method with VAE, and without VAE (namely "Direct Clustering") are shown in Table 2.

Table 2: Evaluation of different combination of clustering. The higher score the better.

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**

As the table demonstrates the results of the approaches, the VAE + OD + AC (extracting outliers, 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. For visualization and interpretability, we use spider plots to visualize the per-cluster mean of the attributes. This makes it easier to interpret the attributes that the clusters were divided based on. 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.

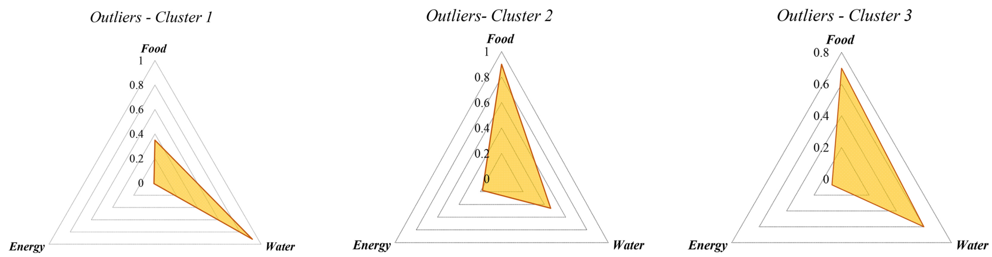

Figure 1: Spider plots for the three example clusters in the outliers group

4 Conclusion

The proposed global clustering approach for cities using ML techniques has two practical implications. First, it will provide space for a comprehensive assessment of cities globally and help identify the aggregate contribution of urban areas to global climate challenges. Second, 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. This assessment for resource use in all cities globally has not been done before, and we believe it will pave the way for a better understanding of opportunities for resource efficiency globally and aid better policy design. The goal for future work is the investigation and identification of 'track shifting' mechanisms and policy interventions that could facilitate urban sustainability.

References

[1] Michele Acuto and Benjamin Leffel. Understanding the global ecosystem of city networks. *Urban Studies*, 58(9):1758–1774, 2021.

[2] Harriet Bulkeley, Michele M Betsill, Daniel Compagnon, Thomas Hale, Matthew J Hoffmann, Peter Newell, and Matthew Paterson. Transnational governance: charting new directions post-paris, 2018.

[3] Jeroen Van der Heijden. Cities and sub-national governance: High ambitions, innovative instruments and polycentric collaborations. 2018.

[4] UN Desa. Revision of world urbanization prospects. *UN Department of Economic and Social Affairs*, 16, 2018.

[5] Jennifer Robinson. Cities in a world of cities: The comparative gesture. *International journal of urban and regional research*, 35(1):1–23, 2011.

[6] Christopher D Elvidge, Kimberly E Baugh, Sharolyn J Anderson, Paul C Sutton, and Tilottama Ghosh. The night light development index (nldi): a spatially explicit measure of human development from satellite data. *Social Geography*, 7(1):23–35, 2012.

[7] Hongwei Xiao, Zhongyu Ma, Zhifu Mi, John Kelsey, Jiali Zheng, Weihua Yin, and Min Yan. Spatio-temporal simulation of energy consumption in china's provinces based on satellite night-time light data. *Applied Energy*, 231:1070–1078, 2018.

[8] Isabel M Horta and James Keirstead. Downscaling aggregate urban metabolism accounts to local districts. *Journal of Industrial Ecology*, 21(2):294–306, 2017.

[9] Liangxin Fan, Lingtong Gai, Yan Tong, and Ruihua Li. Urban water consumption and its influencing factors in china: Evidence from 286 cities. *Journal of Cleaner Production*, 166:124–133, 2017.

[10] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[11] M Schiavina, A Moreno-Monroy, L Maffenini, and P Veneri. Ghsl-oecd functional urban areas. Technical report, JRC Technical Report, 2019.

[12] Amy McNally et al. Fldas noah land surface model l4 global monthly 0.1×0.1 degree (merra-2 and chirps). *Atmos. Compos. Water Energy Cycles Clim. Var*, 2018.

[13] Zhengming Wan, Simon Hook, and Glynn Hulley. Mod11a2 modis/terra land surface temperature/emissivity 8-day l3 global 1km sin grid v006. *Nasa Eosdis Land Processes Daac*, 10(10.5067), 2015.

[14] M Adamovic. Methodology and motivation for numbeo. numbeo. *Inc., Belgrade, Serbia (www. numbeo. com)*, 2021.

[15] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[16] Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*, 2011.

[17] Sidra Arshad, Shougeng Hu, and Badar Nadeem Ashraf. Zipf's law and city size distribution: A survey of the literature and future research agenda. *Physica A: Statistical mechanics and its applications*, 492:75–92, 2018.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] Tadeusz Calinski. A dendrite method for cluster analysis. *Communication in statistics*, 3:1–27, 1974.

[20] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[21] Artessa Niccola D Saldivar-Sali. *A global typology of cities: Classification tree analysis of urban resource consumption*. PhD thesis, Massachusetts Institute of Technology, 2010.

[22] RI McDonald. City water map (version 2.2). *Knowledge Network for Biocomplexity*, 2016.

[23] Euromonitor International. Passport consumers. *https://www.euromonitor.com/*, 2015.

[24] Richard G Newell, Yifei Qian, and Daniel Raimi. Global energy outlook 2015. Technical report, national bureau of economic research, 2016.

[25] Alina Zaharia, Maria Claudia Diaconeasa, Laura Brad, Georgiana-Raluca Lădaru, and Corina Ioanăș. Factors influencing energy consumption in the context of sustainable development. *Sustainability*, 11(15):4147, 2019.

[26] Husi Letu, Masanao Hara, Hiroshi Yagi, Kazuhiro Naoki, Gegen Tana, Fumihiko Nishio, and Okada Shuhei. Estimating energy consumption from night-time dmps/ols imagery after correcting for saturation effects. *International journal of remote sensing*, 31(16):4443–4458, 2010.

[27] Paul Guinness and Brenda Walpole. *Environmental systems and societies for the IB diploma coursebook*. Cambridge University Press, 2015.

[28] Lenka Slavíková, Vítězslav Malỳ, Michael Rost, Lubomír Petružela, and Ondřej Vojáček. Impacts of climate variables on residential water consumption in the czech republic. *Water Resources Management*, 27(2):365–379, 2013.

[29] Rosa Duarte, Vicente Pinilla, and Ana Serrano. Is there an environmental kuznets curve for water use? a panel smooth transition regression approach. *Economic Modelling*, 31:518–527, 2013.

[30] Geoffrey J Syme, Quanxi Shao, Murni Po, and Eddy Campbell. Predicting and understanding home garden water use. *Landscape and Urban Planning*, 68(1):121–128, 2004.

[31] Kirsten Davies, Corinna Doolan, Robin Van Den Honert, and Rose Shi. Water-saving impacts of smart meter technology: An empirical 5 year, whole-of-community study in sydney, australia. *Water Resources Research*, 50(9):7348–7358, 2014.

[32] Bradley Jorgensen, Michelle Graymore, and Kevin O'Toole. Household water use behavior: An integrated model. *Journal of environmental management*, 91(1):227–236, 2009.

[33] Dianne Neumark-Sztainer, Mary Story, Cheryl Perry, and Mary Anne Casey. Factors influencing food choices of adolescents: findings from focus-group discussions with adolescents. *Journal of the American dietetic association*, 99(8):929–937, 1999.

[34] Jan-Benedict EM Steenkamp. Food consumption behavior. *ACR European Advances*, 1993.

[35] Molly C Bernhard, Peng Li, David B Allison, and Julia M Gohlke. Warm ambient temperature decreases food intake in a simulated office setting: a pilot randomized controlled trial. *Frontiers in nutrition*, 2:20, 2015.

[36] The Global Economy. Economic data. *https://www.theglobaleconomy.com/*, 2022.

[37] X Jin and J Han. K-medoids clustering, encyclopedia of machine learning, 2010.

[38] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996.

[39] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Appendix A Data and Model Selection

A.1 Data

In this paper, we developed a machine learning approach to predict energy, water, and food consumption for a total of 9,000 cities around the world. That resulted in three models, each specialized in predicting one resource component using particular sets of features. We adopted a novel approach that relied on different data sources like city night light radiance from satellite images, population, land area, and others, and we used these variables as a proxy to estimate energy, water, and food consumption models. For each model, we experimented and expanded on similar work that was done before in literature, such as choosing variables, using ML models and proportion (non ML) models, and evaluated them using the benchmarks we designed.

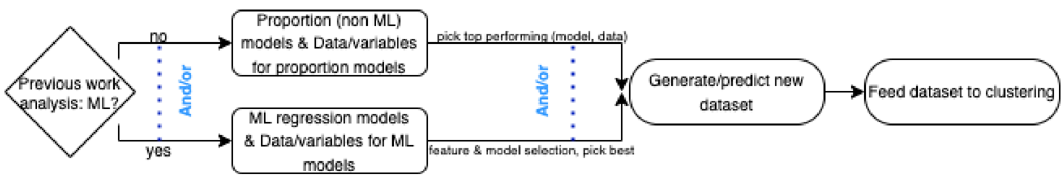

Figure 2: Model & feature selection decision flow

Previous energy, food, and water consumption data per city - for a comprehensive 9000 cities around the world - simply does not exist. Thus, we collected resource data for a subset of cities around the world and processed each dataset to construct our predictive models. We used [21] to obtain energy use values for 155 cities, and [22] to obtain water use for 172 cities. For food, we used [23]. Since the data and context varied greatly between the three resources, we treated each case separately and designed our predictive models accordingly.

A.2 Energy

Data: Countries report their energy consumption values each year [24]. However, little information is reported on the city-level, especially for the units outside the OECD region. We used a random sample of 157 cities [21] that included the average electricity use per capita in kWh for the year of 2012.

Feature Selection and Prepossessing: There are many variables that intuitively correlate with energy consumption such as the available income resources and the location of the city. Horta and Keirstea [8] attempted downscaling the energy consumption from the country to city-level before, but they performed their analysis on London and near-London cities only. Yet, they mentioned several approaches to tackle this challenge such as fitting population-proportion or regression models. We gathered information based on the literature of what affects the energy consumption, and we found out that population, GDP per capita, green house gas emissions, and temperature, [25] were the biggest four indicators of energy consumption [6], [7]. Therefore, we tried to gather as much information on these variables as possible in one universal dataset to perform our analysis. The list of the final variables included: population, population density, area, percentage of area covered by water, percentage of area covered by land, average temperature, city-light radiance, and some vegetation indices. Our methodology was to look for variables that are available for the 9000 cities, so we can generalize our findings.

The selection criteria for these variables are based on the following steps

1. Try different statistical feature selection methods (namely, forward selection, backward selection, and best subset), manually look for features that are consistently top performing across different those feature selection ways, pick them as our best subset.
2. Out of this best subset of features, pick those who can be used for 9000 cities (i.e. if we have a value for each city in our city list of 9000 cities).

Model Selection: Inspired by the work done in [8], we replicated the proposed methods; namely a linear regression with population model, and a multivariate regression with population and area. We benchmarked them with a proportion model we designed using city-light radiance based on a previous study by Letu et al[26]. We created that model by estimating the proportion of the city light radiance of the city to its entire country then multiplied by the total consumption for the country. This method estimated energy consumption values for an entire city. Moreover, we wanted to replicate this approach on the capita-level to see if our assumptions hold.

Chosen Model: City-Light Proportion per capita As tables 3, 4 suggest, the city-light proportion model performed better than the regression ones on the city-scale and the capita-scale across all our statistical metrics presented in B.1. Tables 3, 4 portray our results. The scores are the average of the metric on 10 different random test sets, and each evaluation on one test set was done via 10-cross validation

Table 3: Model Performance evaluated on a test set on energy consumption per capita

Model Benchmark Energy per Capita			
Model/Metric	MAPE	R^2	Ratio Score
Linear Regression with city light and land	193.8%	0.07	234%
Linear Regression with population and land	212%	0.05	264.8%
Population Proportion Model	**67.7%**	**0.77**	**89.5%**

Table 4: Model Performance evaluated on a test set on energy consumption per city

Model Benchmark Energy City-Level			
Model/Metric	MAPE	R^2	Ratio Score
Linear Regression with city light and land	58.9%	0.73	89.7%
Linear Regression with land and population	99.9%	0.03	634.5%
Population Proportion Model	**67.7%**	**0.52**	**89.5%**

A note on the results: While the results we obtained for energy consumption are not relative high (compared to food and water), they are consistent. Our proportion model suggests a linear correlation between our estimation and the ground truth value, albeit a relatively high error rate (MAPE, ratio score). The main challenge was energy consumption was the very little data points we had (157 data points), which restricted our use of more advanced ML.

A.3 Water

Data: Since water consumption data per city for a broad range such as 9000 is not available, we used Urban Household Water Consumption Data [22], which provided water consumption data (liters per capita per day) for years 2014 and 2015. The data corresponded to 289 data points, of which 119 were for 2014 and 170 for 2016. Out of all 289 data points, there were 172 unique cities.

Feature Selection and Prepossessing: Previous studies have suggested various drivers of water consumption. Domestic water use is highly complex and diverse because it can be affected by many factors. For example, one view is that water consumption is highly affected by population: with increasing city population, global water consumption in cities has increased by approximately six-fold, which was twice the rate of population growth [27]. Other views are climate and meteorology [28], socio-demographic profiles [29], household characteristics [30], water availability and conservation [31], and pricing and policies [32]. To determine the right subset of features for our water consumption model, we generated a comprehensive dataset by combining the water consumption data we had with the Euromonitor data [23], which provided us with 62 different features, which span across domains such as socioeconomic, meteorological factors, water supply, etc. The water consumption data we used and Euromonitor data did not align perfectly (i.e. some cities were in one dataset but not the other, and vice versa), thus we had to eliminate cities with no information. The finalized dataset included 172 datapoints across 62 features. To our understanding, using proportion models to estimate water consumption was not supported in the literature review, thus we focused on machine learning models.

Our feature selection process included two steps:

1. Try different statistical feature selection methods (e.g. Recursive Feature Elimination with random forest regressor), manually look for features that are consistently top performing across different those feature selection ways, pick them as our best subset.
2. Out of this best subset of features, pick those who can be used for 9000 cities (i.e. if we have a value for each city in our city list of 9000 cities)

Our methodology is summarized as following: after trying different types of feature selection processes (namely, wrapper: Recursive Feature Elimination with different types of estimators, embedded: lasso regression) we picked a subset that consisted of precipitation, GDP per capita, death rate, land area, population growth, water price, temperature, birth rates, Consumer Price Index, total population.

Among those, we selected only those which we have information for 9000 cities. For example, death rates was consistently found to be a strong predictor for water consumption, but many cities around the world do not provide that data. Our finalized subset of features included precipitation (2015), land temperature (two meters above the ground, for 2015), land area (2015), total population (2015), and water price (in retail stores, a value that was similar between cities within the same country).

This dataset was normalized, since city values, naturally, have a wide range. Outliers in these scenarios are meaningful and thus were not excluded. The water consumption distribution was right skewed as shown in figure 3, thus the models were evaluated on a log scale of the water consumption.

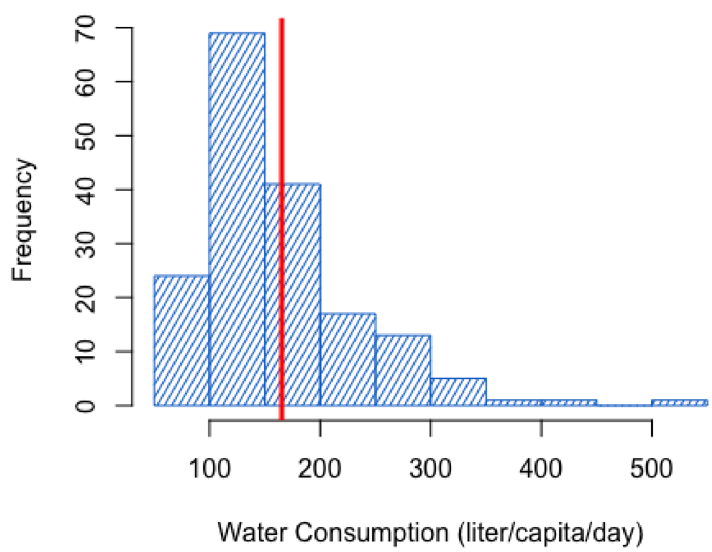

Figure 3: Water Consumption Distribution. The vertical red line shows the mean, which is 165.36

Model Selection: Post picking the best subset of features to estimate water consumption, we tried numerous machine learning models. The main challenge was the lack of data (total 172 points), which had two consequences: first, it disabled us from using large deep learning models that are data hungry. Second, it forced us to focus on a search space of "simple" (classical ML) models with low variance (better generalization), even at the expense of high bias (error). We split the data into train

(80%) and test (20%). We trained the various models on the train data and evaluated it on the test data using k cross validation where k = 10. Since we dealt with so little data, we wanted to control for the case that a random selection of cities for the train and test split does affect the performance (i.e. it may be the chance that a specific random split generated great performance while a different random one did not). To control for that, we evaluated each model on 10 different data splits (i.e. changing the seed number). The results are presented in table 5. We used a baseline model for each model we tested.

Table 5: Model Performance evaluated on a test set. The scores are the average of the metric on 10 different random test sets, and each evaluation on one test set was done via 10-cross validation

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%

Chosen Model: Extremely Randomized Trees (ERT): 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.

A.4 Food

Data: To estimate the food consumption for the 9000 cities, we used Euronmonitor International data [23]. It has information on 1220 cities worldwide, which seemed to be an adequate sample to investigate. The information is available for many years, for consistency sake, we used values for 2015.

Feature Selection and Prepossessing: We relied heavily on the literature to identify which factors influence food behaviors. Some studies [33][34] revealed that economic, social, and physical factors are the major determinant of food consumption for individuals. Another study also pointed out the location and temperature affects people's appetite. [35]. Thus, we tried looking for data that include information on these variables at Euromonitor and Global Economy [36]. Some elements on the list included: housing expenditure, communication expenditure, health-related expenditure, average household number, birth rate, inflation, and growth rate. The feature selection method for food was similar to energy and food: we checked the statistical significance of the the variables we had in our dataset and looked for highest performing set of variables. Additionally, we picked the subset of features that can be applied to the 9000 cities.

Model Selection: Since the food model was developed simultaneously with the water model, we first attempted to take similar regression and proportion strategies. We tried regression models and proportion models based on our energy estimations and city light as described in the energy section A. The regression model results were relatively satisfying. However, we wanted a more robust approach to food. We relied heavily on refining our definition of food consumption. Do we think of it as the expenditure on food and/or food supplies? Intake of protein? Intake of fats? Intake of calories? We decided to define food consumption as the average daily consumption of calories.

Chosen Model: Population Proportion per capita We constructed several models using data from the Global Economy [36] employing each definition of food consumption, and the best one that had the highest R-squared, lowest MAPE, and lowest score in ratio test was the population proportion model of average intake of calories as shown in these results presented in tables 6, 7. The scores are the average of the metric on 10 different random test sets, and each evaluation on one test set was done via 10-cross validation

Table 6: Model Performance evaluated on a test set for food consumption per capita

Model Benchmark Food per Capita			
Model/Metric	MAPE	R^2	Ratio Score
Linear Regression with population	44.1%	0.06	63.8%
Linear Regression with population and land	44%	0.062	61.6%
Population Proportion Model	26.5%	0.681	31%
Linear Regression Population Proportion Model	**22.5%**	**0.711**	**30.2%**

Table 7: Model Performance evaluated on a test set for food consumption per city

Model Benchmark Food City-Level			
Model/Metric	MAPE	R^2	Ratio Score
Linear Regression with population	51.3%	0.55	131.9%
Linear Regression with population and land	79%	0.67	138.5%
Population Proportion Model	**26.5%**	**0.95**	**31%**

Appendix B Evaluation Criteria for the novel Resource Dataset

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 new, we wanted to benchmark it through three different approaches: standard regression metrics (MAPE and R^2 on the out of sample set), statistical characteristics (comparing mean, median, range, variance between the original data and the predicted data), and a Ratio Score metric 1 which we have developed specifically for this task.

Out of these three evaluation metrics, standard regression metrics and Ratio Score are presented in table 1.

B.1 Standard Regression Metrics

We evaluate our model across all counties in the test year with data. We use two standard regression metrics: MAPE and R^2.

MAPE (Mean Absolute Percentage Error) is commonly used as a loss function for regression problems and in model evaluation, and it has a very intuitive interpretation in terms of relative error, which is why we chose it over other similar metrics such as RMSE and MAE which are less intuitive. It is the sum of the individual absolute errors divided by the true values:

$$MAPE = \frac{1}{N}\sum_{i\in\mathcal{T}}\frac{|y_i - p_i|}{y_i}$$

Where N is total number for cities in the test set $\mathcal{T}$, y_i is the true value for city i and p_i is the predicted value for city i.

R^2 is a measure of how much the variation in the data can be explained by the model predictions. Formally,

$$R^2 = 1 - \frac{\sum_{i\in\mathcal{T}}(y_i - p_i)^2}{\sum_{i\in\mathcal{T}}(y_i - \overline{y})^2}$$

where $\overline{y}$ is the average value across the entire test set $\mathcal{T}$. The top of the fraction corresponds to the sum of the squared residuals (RSS: difference between true yield and model prediction). The bottom is the total sum of squares (TSS: the difference between the true value and the average value across the test set), which is proportional to the overall variance of the test set.

B.2 Statistical Characteristics Benchmark

The statistical characteristics of the original data we possessed and our novel data is presented in table 8. We observe that the mean and median are similar, but the range and standard deviation is wider in the original dataset. We believe this happens because the models were trained on very few data points, while making predictions for a many more data points. This motivates the importance of picking the right predictive model based on how well it generalizes on the out of sample data set.

B.3 Ratio Score Metric

The Ratio Score metric we have developed was created to depict how well each predictive model captures the ratio between resource consumption of cities. Table 9 shows an example that was made to show the Ratio Score between two cities for reference. When we used Ratio Score to evaluate our models, we used the average Ratio Score on the out of sample dataset. We define a good Ratio Score to be below 30%. Ratio Score metric is presented in algorithm 1. In words, the Ratio Score Metric for city i calculates the true and predicted ratio between city i in the test set and all the other cities

Table 8: Data set benchmark via 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

(numbered $1...N$, if the test set includes N cities) in that set except for city i. Then, it calculates the MAPE between the true and the predicted Ratio Scores, and averages through all cities.

Table 9: Ratio Score for water consumption between two cities. The MAPE is calculated by $\frac{|true-predicted|}{true}$, for example $\frac{|171-188.92|}{171} = 10.48\%$. Ratio is calculated by dividing the values between cities, for example, true ratio here is $\frac{171}{220} = 0.78$

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

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 $\mathcal{R}$

B.4 Limitations:

Each model has its limitations based on the dataset available. The energy model doesn't accurately represent super-mega cities like Texas, NY, and LA since the light radiance doesn't go beyond a specific point. That is, a city will use more energy, but it is bright enough that more consumption doesn't get captured in the satellite images. For water, since we worked with very few data points (172 total, 138 used for training), we need to make sure the distribution of cities is generalizable enough; namely, it represents a comprehensive set of countries and cities to match (or at least be close to) the 9000 cities we used to predict. This, unfortunately, is not under our control since water consumption data is very scarce. Regarding food, the consumption of calories doesn't capture the individual variations of the type of consumed food, i.e. healthy or junk. We assume that the only variation in the city consumption is their population and the food sources available.

Appendix C Clustering

We experimented other standard clustering techniques such as: K-Means algorithm (KM) [37] and Gaussian Mixture Models (EM) [38] and compared the results of the combination with the Outlier Detection (OD) and Variational Autoencoders (VAE) and reported the results, seen in table 10. As seen in the paper, our proposed three-fold clustering of (1) outlier detection (2) VAE (3) agglomerative clustering performs best, both in terms of CHI and SC. Figure 4 shows the spatial distribution for 1100 cities identified in the outliers group over five clusters. It can be noted that the outliers group mostly included major cities in the developed North, parts of the United States, West-Northern Europe, and major cities in Asia.

Table 10: Silhouette Coefficient outputs and Calinski-Harabasz Index for the tested methods

	Calinski-Harabasz Index		Silhouette Coefficient	
Algorithm/Method	with VAE	Direct Clustering	with VAE	Direct Clustering
K Means	39851.71	5031.95	0.315	0.368
Agglomerative	**42495.66**	4345.80	**0.466**	0.452
Gaussian Mixture	8190.34	1312.97	0.206	0.0252

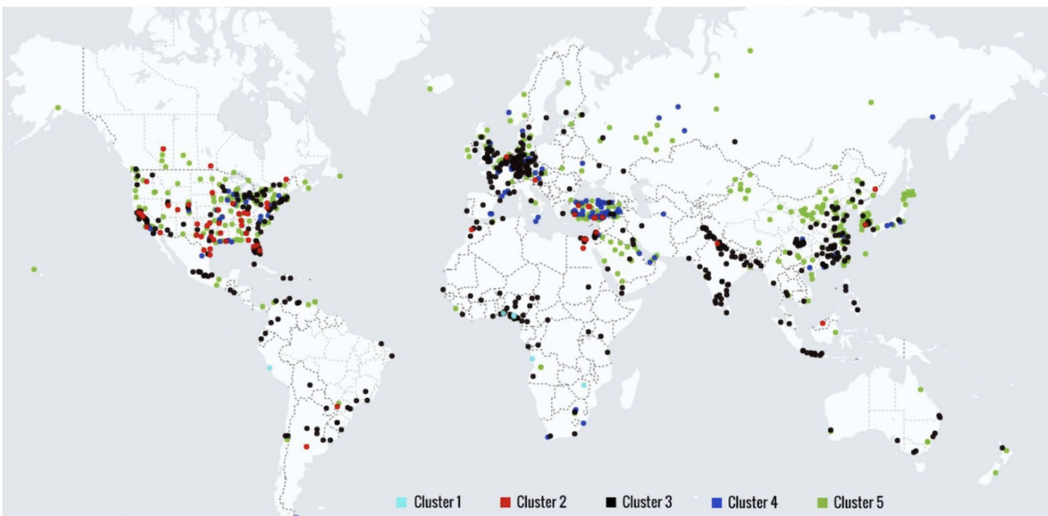

Figure 4: Spatial distribution of the outliers group

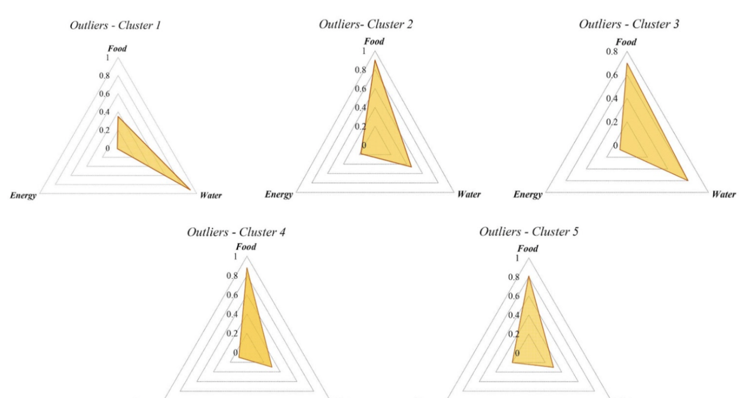

Figure 5: Spider plots for the five main clusters in the outliers group

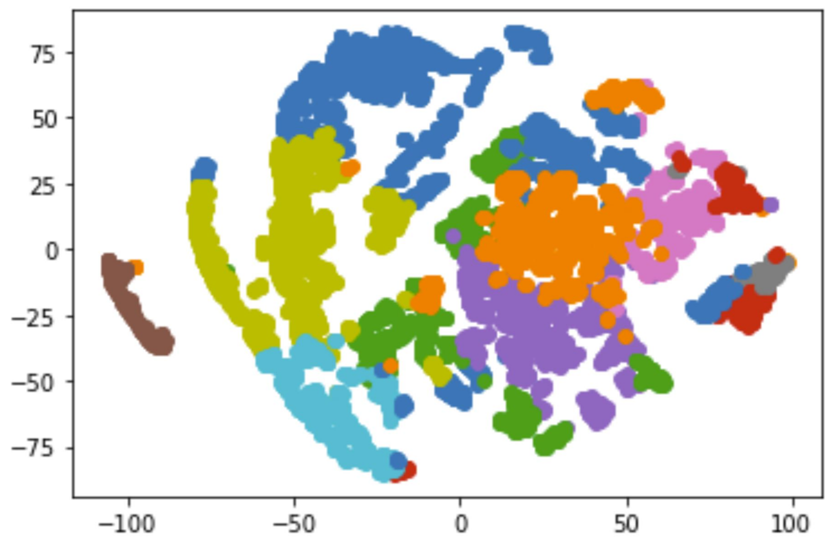

Figure 6: t-SNE using AC

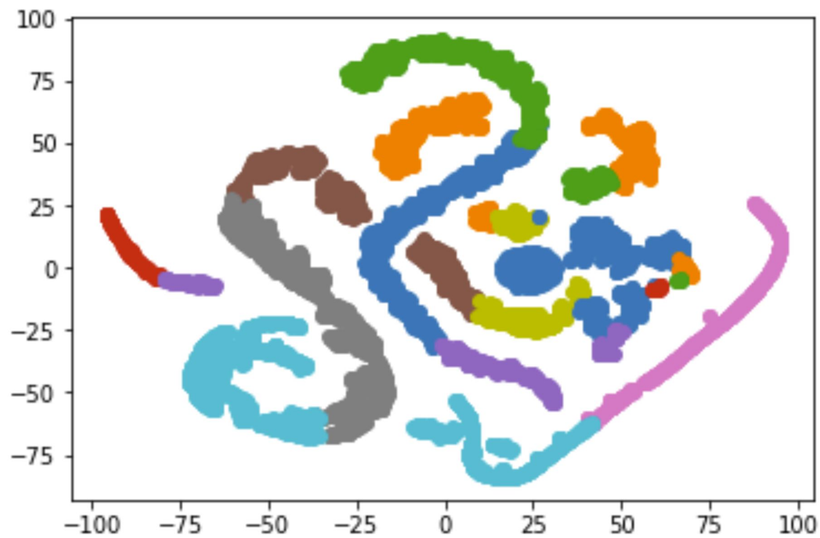

Figure 7: t-SNE using OD+VAE+AC

In addition, we provided t-SNE [39] plots to visually analyze the performance of the clustering algorithms we tested. t-SNE is a visualizing algorithm that visualizes multi-dimensional data by projecting them to a 2D space. This is the 2D representation of the t-SNE algorithm. As shown in figures 7, 6 and aligned with our quantitative metrics (Calinski-Harabasz Index and Silhouette Coefficien), VAE clustering has much clearer, and more separable clusters, suggesting it performs better than an alternative, baseline clustering approach.