Poster: A Realistic Dataset Generator for Smart Grid Ecosystems with Electric Vehicles

Georgios Charalambidis
School of Electrical & Computer Engineering
Technical University of Crete, Greece
gcharalampidis@isc.tuc.gr

Charilaos Akasiadis
Institute of Informatics & Telecommunications
NCSR “Demokritos”, Greece
cakasiadis@iit.demokritos.gr

Emmanouil S. Rigas
School of Medicine
Aristotle University of Thessaloniki, Greece
erigas@auth.gr

Georgios Chalkiadakis
School of Electrical & Computer Engineering
Technical University of Crete, Greece
gehalk@intelligence.tuc.gr

ABSTRACT

Research on the deployment and employment of electric vehicles (EVs) in the emerging Smart Grid, typically requires access to large datasets containing data that is rich and reliable. Such datasets are hard to come by in the wild due to various privacy and sensitivity considerations. In this paper, we design a dataset generator for large-scale EVs charging management. The generator *(i)* takes as input anonymized real-world datasets describing different energy generation and demand types, as well as charging profiles of EVs and corresponding trip and type information; *(ii)* fits a variety of machine learning models using this data as training sets; and *(iii)* generates new synthetic data that adheres to the same principles and relationships as the input. The generator comes complete with data smoothing and dataset summarization, visualization, and comparison abilities that users can utilize via a web-based interface; and is offered as a free-to-use tool to the research community.

CCS CONCEPTS

• **Computing methodologies** → *Data assimilation.*

KEYWORDS

Dataset Generator, Electric Vehicles, Charging, Smart Grid

ACM Reference Format:
Georgios Charalambidis, Charilaos Akasiadis, Emmanouil S. Rigas, and Georgios Chalkiadakis. 2022. Poster: A Realistic Dataset Generator for Smart Grid Ecosystems with Electric Vehicles. In *The Thirteenth ACM International Conference on Future Energy Systems (e-Energy ’22), June 28–July 1, 2022, Virtual Event, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3538637.3538755

1 INTRODUCTION

Although EV deployment has been growing rapidly over the past ten years, many challenges still arise in different levels. For instance, the charging infrastructure needs to be properly placed to service large numbers of customers [10]; a variety of critical elements of vehicle-to-grid (V2G) economics have to be accurately identified and confronted [9]; while there is a need for the appropriate redesign of the electricity distribution network to accomodate EVs [2] As such, there is an utmost need for reliable data for a variety of purposes—e.g., for understanding behaviors, exploring flexibility, and extrapolating results for other similar cases or sites. However, the lack of reliable data to advance this highly interdisciplinary research, is a known problem in this rising market [8]. Even when datasets are publicly available [1, 7], limitations still arise—e.g., the range/size of the datasets is limited, or the data available may be subject to copyright and not freely shared for use.

This calls for simulators to address data engineering needs and enable the design and analysis of novel components with minimum risk. Lee et al. [5] introduced ACN-Data, a dynamically populated dataset of EVs charging in workplaces, which, however, only includes workplace EV charging data, but no trips-related or production and consumption data. RAMP-mobility [6], relies on user input that indicates specific properties of the data and generates mobility and demand data for a number of EVs, using a stochastic model. EMOBPY [3], employs relative frequencies calculated using real data to generate time series of vehicle mobility, energy consumption, connectivity, and electricity demand of charging tasks.

Against this background, we show how to incorporate several anonymized publicly available electricity production and consumption datasets into a novel generation framework for Smart Grid ecosystems with EVs. Our framework can produce new, synthetic data governed by the same principles as the original, but does not compromise privacy and thus can be made available to the public.

2 METHOD OVERVIEW

The original collected data (from any available source) and the generated data is divided into two main categories: *(a) energy production and consumption* and *(b) EVs’ and drivers’ behavior data.* The first includes data related to the total consumption of renewable and non-renewable energy for a given region’s infrastructure and the respective production. The second refers to EVs and their characteristics; the specifications of the available EV chargers; and to drivers’ behavior, regarding the trips they perform, and the number of charging sessions and the time of the day these take place.

e-Energy ’22, June 28–July 1, 2022, Virtual Event, USA

ACM ISBN 978-1-4503-9397-3/22/06…$15.00
https://doi.org/10.1145/3538637.3538755

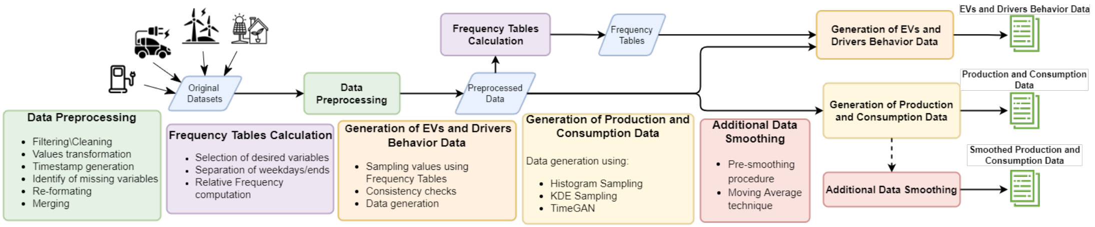
Original Datasets
Data Preprocessing
Preprocessed Data
Frequency Tables Calculation
Frequency Tables
Generation of EVs and Drivers Behavior Data
EVs and Drivers Behavior Data
Generation of Production and Consumption Data
Production and Consumption Data
Additional Data Smoothing
Smoothed Production and Consumption Data
Data Preprocessing
Filtering\Cleaning
Values transformation
Timestamp generation
Identify of missing variables
Re-formating
Merging
Frequency Tables Calculation
Selection of desired variables
Separation of weekdays/ends
Relative Frequency computation
Generation of EVs and Drivers Behavior Data
Sampling values using Frequency Tables
Consistency checks
Data generation
Generation of Production and Consumption Data
Data generation using:
Histogram Sampling
KDE Sampling
TimeGAN
Additional Data Smoothing
Pre-smoothing procedure
Moving Average technique

Figure 1: Schematic overview of workflow components in our system

Figure 1 shows the workflow and the components of our system. First, data is preprocessed to generate compatible timestamps, consistent across all generated files, as the original data comes from different sources and has different formats. In addition, irrelevant and/or incomplete data is omitted, while some other is transformed to facilitate our work. We also extract new information—necessary for the data generation—by combining and merging data and files. In the case of energy production and consumption, the preprocessed dataset is fed directly to our proposed generation methods: histogram (HIST) and kernel density estimation (KDE) sampling [4], and time-series generative adversarial networks (TimeGAN) [11]. For EVs’ and drivers’ behavior, computing the frequency tables of specific variables of interest (e.g., travel speed, time of connection, etc.) is also required. The calculated frequencies, along with specific constraint checks—in order to create realistic data and as close to the original—are used to generate the corresponding dataset. Finally, additional smoothing can be applied in the output.

User Configurations. To generate data, the user can either *(a)* use a web-based GUI that eases the configuration of the generator, or *(b)* execute generation script templates and set values to required configuration parameters via JSON files and the command line. For the latter, the user must specify: the number of EVs; the time horizon length; the desired data generation technique among the available ones; the categories of EVs to be included in the generated dataset; the types of EV chargers to be considered; the additional smoothing parameters, if desired. This information contains all required configurations for initiating the generation process. The generated dataset comes in the form of CSV files, together with summarizing statistic tables and figures—specifically histograms, bar plots and time-series curves. Furthermore, any number of original datasets—different from the default included in our work—can also be incorporated for the training of the generation models. This is particularly helpful for organizations that need to protect private data, but are still in need to share anonymized information with third parties. The web-based GUI uses a backend web server implemented in Flask—a Python micro web framework—while the frontend part was implemented using pure *HTML* and *JavaScript*.

3 RESULTS OVERVIEW

We compare the distributions of the original data considered as input, to those of the synthetic resulting from the generation process. Intuitively, the difference should not be too large so as to imply different underlying models. The fitting of the models and the generation of the synthetic production and consumption datasets for a time horizon of 4 years including 100 EVs, takes 7 minutes for the HIST method, while KDE requires 3 to 4 hours, and TimeGAN about 6 hours for data generation, on a computer with an Intel Core i7-5500U CPU 2.40GHz processor and 8GB RAM. For the EV and Drivers Behavior, the Frequency Tables method required 30 minutes. To quantify a “distance” between the distributions under comparison, we employ the well-known Kolmogorov - Smirnov D statistic. KDE and HIST sampling methods’ results are very similar, and both outperform TimeGAN by a relatively large margin. More specifically, the value of the K-S statistic for KDE and HIST for all columns is really small, implying the produced data distribution is a good fit of the original. Indeed, when we create a relatively large number of data using KDE and HIST, the overall picture of the generated data is very close to the original, both in terms of distributions, and in time-series forms. However, for specific types of data with no periodicity, the methods tested are less effective in producing similar time series patterns to those of the original data.

In the future, we plan to evaluate the application of additional data analysis techniques for datasets that exhibit some periodicity in their values; and at the same time, to explore ways to more accurately simulate data that does not appear to be periodic. Focus will also be put on the positioning of the EVs towards generating more realistic mobility-related information.

REFERENCES

[1] S. Barker, A. Mishra, D. Irwin, E. Cecchet, P. Shenoy, J Albrecht, et al. 2012. Smart*: An open data set and tools for enabling research in sustainable homes. *SustKDD, August* 111, 112 (2012), 108.

[2] M. G. Flammini, G. Prettico, G. Fulli, E. Bompard, and G. Chicco. 2017. Interaction of consumers, photovoltaic systems and electric vehicle energy demand in a Reference Network Model. In *2017 Int. Conf. of Electrical and Electronic Technologies for Automotive*. IEEE, 1–5.

[3] C. Gaete-Morales, H. Kramer, W.-P. Schill, and A. Zerrahn. 2021. An open tool for creating battery-electric vehicle time series from empirical data, emobpy. *Scientific data* 8, 1 (2021), 1–18.

[4] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. 2004. *Nonparametric and semiparametric models*. Vol. 1. Springer.

[5] Z. J. Lee, T. Li, and S. H. Low. 2019. ACN-Data: Analysis and applications of an open EV charging dataset. In *Proc. of the Tenth ACM Int. Conf. on Future Energy Systems*. 139–149.

[6] A. Mangipinto, F. Lombardi, F. D. Sanvito, M. Pavičević, S. Quoilin, and E. Colombo. 2022. Impact of mass-scale deployment of electric vehicles and benefits of smart charging across all European countries. *Applied Energy* 312 (2022).

[7] Pecan Street 2021. *Pecan Street Dataport*. Retrieved February 24, 2021 from https://www.pecanstreet.org/dataport/

[8] D. Pevec, J. Babic, and V. Podobnik. 2019. Electric vehicles: A data science perspective review. *Electronics* 8, 10 (2019), 1190.

[9] D. M. Steward. 2017. *Critical elements of vehicle-to-grid (v2g) economics*. Technical Report. National Renewable Energy Lab.(NREL), Golden, CO (United States).

[10] Y. Xiong, B. An, and S. Kraus. 2021. Electric vehicle charging strategy study and the application on charging station placement. *Autonomous Agents and Multi-Agent Systems* 35, 1 (2021), 1–19.

[11] C. Zhang, S. R. Kuppannagari, R. Kannan, and V. K. Prasanna. 2018. Generative adversarial network for synthetic time series data generation in smart grids. In *2018 IEEE Int. Conf. on Comm., Control, and Comp. Tech. for Smart Grids*. 1–6.