**Final Research Paper**

George Selly

Colorado State University Global

MIS581: Capstone—Business Intelligence and Data Analytics

Dr. Osama Morad

February 4, 2024

Contents

[1.0 Abstract 2](#_Toc158236015)

[2.0 Introduction 3](#_Toc158236016)

[3.0 Objectives 4](#_Toc158236017)

[4.0 Overview of Study 4](#_Toc158236018)

[5.0 Research Question 5](#_Toc158236019)

[6.0 Literature Review 5](#_Toc158236020)

[6.1 Theoretical Framework 5](#_Toc158236021)

[6.2 Preferred Modeling Approaches 6](#_Toc158236022)

[6.3 Power Usage Signal Processing Premises 6](#_Toc158236023)

[7.0 Research Design 8](#_Toc158236024)

[7.1 Methodology 8](#_Toc158236025)

[7.2 Methods 9](#_Toc158236026)

[7.2.1 Data Collection and Pre-Processing 9](#_Toc158236027)

[7.2.2 RStudio Regression Models 10](#_Toc158236028)

[7.2.3 Python Neural Network Model 11](#_Toc158236029)

[7.3 Limitations 12](#_Toc158236030)

[7.4 Ethical Considerations 13](#_Toc158236031)

[8.0 Findings 14](#_Toc158236032)

[8.1 Regression Analysis in RStudio 14](#_Toc158236033)

[8.2 Neural Network Analysis in Python 17](#_Toc158236034)

[9.0 Conclusion 18](#_Toc158236035)

[10.0 Recommendations 19](#_Toc158236036)

[11.0 References 20](#_Toc158236037)

[12.0 Tables and Figures 23](#_Toc158236038)

# 1.0 Abstract

Aggregate power usage modeling is an increasing challenge for utility providers, given both the increasing prevalence of intermittent, weather-dependent power sources and the increasing reliance on batteries for supply/demand rectification vs. peaking stations. To this end, the increasing prevalence of smart devices is cause for optimism. This prevalence enables Internet of Things-based, highly detailed data pipelines, which promise to improve downstream analytics vs. conventional, lower-fidelity data pipelines. However, given the security risks imposed by both the personal identifiability of this data and IoT-based pipelines' reliance on embedded systems, those gains must be proven to justify using IoT-based analytics in a given domain. This quantitative study applies that test to the power usage modeling domain. It investigates whether models of geographic clusters of homes' power usage perform as well as models of households' power usage, while also investigating whether appliance-level breakdowns of power usage improve predictive models. This study found that modeling total usage across these clusters provides optimal results; access to household-level or appliance-level usage data does not improve model outcomes.

**Keywords**: internet of things, power usage modeling, data minimization

# 2.0 Introduction

A key aspect of utility providers’ mission statements—providing uninterrupted access to the grid, for both producers and consumers—demands constant load balancing despite highly volatile conditions (Mamun et. al., 2020). Essentially, providers must be able to anticipate events that threaten grid stability—such as peak residential usage periods and weather events that change renewable energy sources’ outputs—so they can use batteries, peaking stations, etc. to address those events. Addressing these events optimally, such that stability is maximized while operational costs are minimized, requires an accurate prediction algorithm for overall power usage (Mamun et. al., 2020). One tool that promises to improve these algorithms is the Internet of Things: given the widespread adoption of both smart meters and smart home appliances, utility providers can access power usage data at the household level or the appliance level. This high-resolution sensor data can theoretically act as the basis for better residential power usage models, thus yielding improved grid stability and lower electricity costs. However, streaming vast quantities of IoT data into analytics systems vastly increases cloud compute costs**,** while also exposing consumers to significant data privacy risks (Boeckl et. al., 2019). After all, IoT systems increase the amount of personally identifiable data tied to data subjects while transmitting that data via a wide network of embedded systems that can be tampered with (Boeckl et. al., 2019). And while not all this data is explicitly identifiable, it is often deanonymizable due to the existence of unique usage patterns that are traceable over time (Boeckl et. al., 2019). This forces utility providers to determine whether the increased data resolution from IoT data pipelines improves power usage models enough to justify these risks and externalities.

# 3.0 Objectives

Given these data security risks inherent to IoT, risk mitigation requires datasets be aggregated, anonymized, and restricted in their feature sets as much as possible without degrading analytics outputs (Ancell, 2022). The European Union’s GDPR even codifies this best practice into law: it forces companies to prove that their stored data is both necessary and minimized (GDPR Summary, 2018). This general need to justify the storage and transmission of high-definition user data requires power industry-specific research, aiming to determine the most minimal form of power usage data that still provides optimal analytics outputs. This study addresses that need: it models power usage data—both at different granularities and with different model families—at various granularities to determine the data storage granularity which optimally balances analytics costs, data security risks, and analytics quality.

# 4.0 Overview of Study

Per Hosking et. al. (2013), weather and prior usage are generally sufficient to predict power usage graphs. Following this, hourly power usage data on 10,000 simulated North Carolinian households was collected from Thorne, Mortveit, & Marathe (2023), and was cross-referenced with local weather data from the National Centers for Environmental Information (n. d.). This data was modeled at four granularities:

* Power consumption for 1 type of usage (washing machine usage, dishwasher usage, etc.), for 1 house within 1 hour.
* Power consumption overall, for 1 house within 1 hour.
* Power consumption for 1 type of usage, for a 50-house geographic cluster within 1 hour.
* Power consumption overall, for a 50-house geographic cluster within 1 hour.

This modeling will be performed in two ways at each granularity: via multiple regressions in RStudio and via Neural Networks created using Python’s TensorFlow library, seeing as these are common and effective means of modeling power usage in literature (Neo et. al., 2023, Buonanno et. al., 2022). This combination of 4 granularities and 2 model approaches will yield 8 RMS error and accuracy values for hourly average home energy consumption, where the derived accuracies of these models are used to discuss which data granularity is most appropriate for power usage prediction analytics.

# 5.0 Research Question

The practical uncertainties in determining optimal ETL configurations for residential power usage modeling systems can be summarized by the following research question:

Does access to high-granularity residential power usage data—available either on a per-household or per-appliance basis—improve the accuracy of models for aggregate residential power usage vs. partially aggregated power usage data?

# 6.0 Literature Review

## 6.1 Theoretical Framework

A summary of power usage prediction literature is shown in Figure 1 (Buonanno et. al., 2022, Li & Zha, 2018, Hosking et. al., 2013, Neo et. al., 2023, Li & Zha, 2018, Zhu et. al., 2019). The all-important modeling node is highlighted in red, while the research questions are highlighted in blue.

## 6.2 Preferred Modeling Approaches

The literature largely contains successful implementations of two model families—multiple regressions and neural networks—while autoregressive models are infrequently used despite their intuitive applicability (Buonanno et. al., 2023, Neo et. al., 2023). Additionally, Hosking et. al. (2013) show that Kalman filters are useful for short-term energy prediction, though their model is only used to predict the following 15 minutes’ worth of power consumption. This makes Kalman filters useful for addressing utility providers’ immediate load balancing needs, while not being as useful for addressing slightly longer-term prediction needs. Neo et. al. (2023) show that in a case study of Singaporean household power usage prediction, geographically weighted regressions perform better than unweighted regressions; it intuitively follows that data points that are geographically closer to a given point should be more useful for predicting its usage patterns and responses to stimuli. As for neural networks, while vanilla implementations of neural networks show good performance and are often used, LSTMs are generally preferred in literature due to their retention of state information, allowing the model to be trained on many users while processing each household’s power usage graph as a separate time series (Buonanno et. al., 2023). However, they require immense amounts of memory to train and apply to large datasets such as household-level power usage data, making LSTMs more applicable for modeling aggregate specifics than modeling of highly granular data.

## 6.3 Power Usage Signal Processing Premises

Figure 1’s literature summary graph is largely based on four premises contained in the literature, and tries to extend them to determine ETL configuration practices:

1. Household-level power usage data can be disaggregated into per-appliance signals statistically (Li & Zha, 2018). Since most of the variance in power usage comes from a small number of high-power appliances—each of which uses a different amount of power for a different duration—household power signals can be easily broken down into appliance-level signals (Li & Zha, 2018).
2. Per-appliance signals—whether inferred from the overall signal or retrieved from smart device signals—can be used to infer household configurations, residents’ daily schedules, and other personal information (Zhu et. al., 2019).
3. Household-level power usage data patterns are distinctive enough to reidentify most data points’ addresses; Buchmann et. al. (2013) were able to deanonymize 68% of data points in a sample dataset. This allows inferred signal information to be cross-referenced vs. other aspects of identified individuals’ data profiles such as location data.
4. A single model of household-level power usage trained on all points is as accurate as many models trained on each house separately (Buonanno et. al., 2022).

The first three premises imply that household-level data—and especially appliance-level data—are useful enough for data miners to pose a significant cybersecurity risk even if that data is anonymized. The final premise questions the utility of maintaining that identifiable data at such a granular level post-staging for the purpose of power usage prediction; if general patterns are as predictive of a household’s power usage as that household’s own history, then there is little need to preserve data at that household-specific level (Buonanno et. al., 2023). These premises imply that storing or transmitting highly granular power usage data—either at the household level or at the appliance level—imposes grave data security risks while having an uncertain impact on predictive model quality. Buonanno et. al. (2022) also states that globally trained models are more robust and practical to implement across a service area with millions of households: if all else is equal, models for power usage based on general factors are preferred over models that are based mostly on household-specific details. These details discourage the use of high granularity data in power usage prediction analytics: usage of high granularity data in analytics will need to be justified by substantially better results alongside strict data security protocols.

# 7.0 Research Design

## 7.1 Methodology

This study is rooted in the quantitative tradition: though Figure 1’s two highlighted research question sub-concepts are deducible from premises in literature, they will be experimentally tested by interrogating the largely numeric dataset with quantitative, statistically grounded modeling methods (O’Leary, 2021). This methodological approach enables an objective, exploratory study of the research questions with a reasonable scope, while making low-confidence claims about the population-wide applicability of findings (O’Leary, 2021). After all, “does maintaining highly granular power usage data improve power usage model accuracy across a generic population, across all effective models” is a more sweeping statement than “does maintaining highly granular power usage data improve power usage model accuracy across a sample of North Carolinian households, across two model types.” The former, more general statement can be hypothetically proved by deriving a robust theoretical basis for information gain in power usage system modeling—which would allow one to demonstrate which types of data disaggregation yield useful information gain—while it can also be proved using a series of quantitative experiments conducted using different datasets and models, where each corroborating study improves the confidence of that theory. This latter, decentralized truth-gathering method acts as the broader context and justification for each quantitative experiment such as this one (O’Leary, 2021).

## 7.2 Methods

### 7.2.1 Data Collection and Pre-Processing

High-resolution usage data covering 10,000 North Carolinian households’ power usage levels was provided by Thorne, Mortveit, & Marathe (2022). This data exists on an hourly basis and covers all of 2014. The files can be accessed by using HTTP requests to descend through NET.Science folders based on their metadata, per instructions provided by D. Machi (personal communication, January 5, 2024). Alongside usage data, weather data for 20 local weather stations from the National Centers for Environmental Information (n. d.) was retrieved. The coordinates of each weather station should be adequate to approximate local weather conditions between stations. Of course, given privacy concerns, the usage data provided by Thorne, Mortveit, & Marathe (2022) does not provide household coordinates, so one cannot use precise home coordinates to find the closest weather station for each usage row (National Centers for Environmental Information, n. d.). However, one can generally approximate a home’s closest weather station by finding the closest weather station to that home’s county’s geographic center (United States Census Bureau, 2023a). In this way, the state of North Carolina can be divided into 20 regions as shown in Figure 2, where each region corresponds to a separate weather station whose weather data will be joined with usage data for homes in the region (DIY Maps, n. d., National Centers for Environmental Information, n. d., United States Census Bureau, 2023a).

Using the attached usage\_data\_extractor.py, usage data was downloaded and joined with weather data. Essentially, the code performed a left join between usage data and weather data; for each usage data point, its county’s matching weather point is defined as the earliest point among the list of observations that occurred at the closest weather station within the same 1-hour window as the datapoint, if any such points exist. Once this data was joined, rows with any null or labelled-as-suspect weather values were filtered from the dataset (National Centers for Environmental Information, n. d.). This yielded a 7.9 GB, CSV dataset formatted according to the data dictionary shown in Table 1, which was titled “MIS581\_final\_project\_data\_filtered.csv” (D. Machi, personal communication, January 5, 2024, National Centers for Environmental Information, n. d., National Oceanic and Atmospheric Administration, n. d., Thorne, Mortveit, & Marathe, 2023).

### 7.2.2 RStudio Regression Models

In analytics projects generally, it is an asset for one’s first used analytics method to be highly interpretable. This helps one better understand the dataset, which allows follow-up steps (engineering new features, fixing data quality issues, etc.) to be created after minimal time investments. This makes multiple linear regression an ideal first step, especially when paired with one of the various data analytics tools that can handle regressions. And for power usage modeling, regression is a common and accurate modeling method (Buonanno et. al., 2022). This accuracy can be explained by separating users’ power consumption patterns into three main components:

* Baseline, static usage.
* Daily cycling of usage patterns.
* Weather pattern-based disruptions or modifications of usage.

User responses to both daily schedules and weather patterns can be seen as both partially dependent on user-specific quirks and partially dependent on collective behaviors shared by most data points. For all components, though, the responses should be essentially stationary; past behavior for a specific user, weather condition, and time should be replicated if these conditions re-emerge, making direct model dependence on past usage unnecessary.

For this use case, regression models will be created both for each group and for each household: many small models will be created in lieu of one large model that uses group IDs or home IDs as predictive categorical variables. This model splitting saves memory: as Figure 3 shows, applying a multi-thousand-dimensional regression model to a dataset with 59.7 million rows requires an impractically large and sparse weights matrix. This removes the ability to model certain variables’ collective coefficients across the dataset: each home or group will be modelled entirely on its own terms. This increases overfitting risks for per-home modeling, though this can be seen as testing per-home usage models purely on their own terms. After all, averaging the impacts of variables on home power signals for robustness’s sake is what defines statewide or per-group models.

Of course, there are many low-code tools that can perform regressions and return summary statistics: tools that include SAS Studio, SAS Enterprise Miner, and Google BigQuery. All three of these tools were attempted but they were each unsuitable: SAS Studio has a 1 GB dataset limitation, SAS Enterprise Miner freezes when attempting to import the large dataset, and Google BigQuery doesn’t have the same flexibility in analysis that the R programming language has. Between these challenges and the custom nature of the needed regression model splitting, RStudio analysis—performed in the R programming language, without the aid of low-code tools—was chosen for this study.

### 7.2.3 Python Neural Network Model

Here, feed-forward/vanilla neural networks were applied to the dataset using the Keras library in Python. LSTMs were eschewed due to their memory requirements: after all, running a Keras LSTM on ~59 million row dataset, a 32-dimension input vector (hour\_of\_day + 8 numeric variables) and 100-length input sequences would require 189 billion entries, which far exceeds the RAM on the computer used to perform these tests. So, as a more practical, exploratory application of neural networks, feed-forward neural networks were used instead. For a feed-forward neural network, the neural network acts as an input/output function just as regression does: it considers no previous points, and it does not consider user-specific information when predicting points. And to help the neural network generalize, group and home IDs were not used as categorical variables here: weights were entirely trained on collective behavior at the given granularity. As a result, this application of neural networks was expected to yield more underfitted models than any per-user regressions, as in those regression models, coefficients are determined on a per-user basis rather than on a collective basis. Buonanno et. al. (2022) state that neural network accuracy (in their tested case, LSTM NN accuracy) is consistently on par with regression accuracy, so that is also an expectation that carried forward to these tests.

## 7.3 Limitations

This study is limited in five major ways:

* The power usage data is simulated; depending on the accuracy of Thorne, Mortveit & Marathe’s (2023) generative algorithm, model patterns may not line up fully with studies on real data.
* Historic weather data is only available on a per-weather-station basis, which merely approximates the weather conditions at the exact time and place of each power usage data point.
* The study is restricted to North Carolinian energy usage patterns in 2014, which do not necessarily reflect general usage patterns due to cultural and socioeconomic impacts on power usage patterns (Natividad & Benalcazar, 2023).
* The study is restricted to two model types—multiple regressions and neural networks—despite that it addresses a research question that applies to modeling efforts in general.
* This analysis was conducted on a single computer with 32 GB of RAM, effectively limiting the size of datasets/data structures due to hardware crashes.

The first two limitations stem from data sourcing limitations: the first limitation is due to the lack of open-access, high-definition, big datasets on power usage, while the second limitation is due to low spatial resolution for historic weather datasets. The third and fourth limitations are due to limitations on project scope, where these limitations aim to deliver usable answers to defined research questions within specified resources and timelines. And the final limitation is appropriate for exploratory research: while distributed computing/big data processing methods should be applied to this case in the future, working on a single computer is sufficient for an initial investigation.

## 7.4 Ethical Considerations

Given the simulated nature of the dataset, there are no ethical issues associated with handling the data itself; data points cannot be traced back to real people (Thorne, Mortveit, & Marathe, 2022). However, this is aimed at helping utility providers manage security risks associated with storage and transmission of sensitive data: its findings are inherently centered on ethics. And the interpretation of findings section is inherently subjective; if higher data granularity improves model outputs, the utility provider will need to determine whether those improvements justify consumers’ increased exposure to data security risks. Like other domains impacted by IoT, gaps in legislative protections of consumer data give organizations plenty of discretion when deciding how to balance data security with model accuracy: decisions involving ETL pipeline setup, proper employee roles, data handling protocols within analytics departments, and several other factors. Ultimately, the impact of data granularity on consumer data security should be thought of as one ethically charged component within a larger data management lifecycle, where every part of that lifecycle must be aimed at balancing IT costs, data security risks, and the accuracy of its analytics products.

# 8.0 Findings

## 8.1 Regression Analysis in RStudio

All R code used in this section is attached in the file regression\_power\_usage\_analysis.R. As a first step, summary statistics for the dataset were generated and shown in Figure 4. The dataset’s processed form exists at the same granularity as was present in the raw usage data provided by Thorne, Mortveit, & Marathe (2023): one point per home & per hour. That granularity informs some of the distributions observed here. While many of the variables show normal distributions, both precipitation and most of the power variables share common characteristics: they contain many zeros, they are heavily skewed to the right, and the logarithms of these distributions (once zeros are excluded) show normality. For precipitation, this is expected: North Carolina’s subtropical climate yields heavy, quick downpours: frequent sampling will naturally yield a heavily right-skewed distribution. This can be normalized by setting the variable precipitation\_normal to asinh(precipitation / mean(precipitation)): this transformation is robust, accounts for zeros in the signal, and recognizes the lognormal nature of the remainder of points.

As for power usage, the distributions are expected given the per-household nature of the dataset. For any input vector x, power usage in any given category for a household can be seen as the combination of two functions APPLIANCE\_POWER(x) \* APPLIANCE\_ON(x): certain factors determine whether the appliance is turned on or not, while other factors determine the rate of electricity usage when that appliance is turned on. Thus, a logistic regression can be created to predict whether power > 0, and a logarithmic regression can be used to predict the power level among points > 0 (assuming the points > 0 can be modelled logarithmically). To test this, a logarithmic regression was created predicting log(total\_kw | total\_kw > 0) as a function of weather variables & hour of day. If log(power) can be predicted by a regression of these variables, then there should be homoskedasticity: the distribution of residuals (log(real power) – log(predicted power)) should be roughly the same across the input variable space. Figure 5 shows this by providing boxplots of the residuals across each fitted value decile: 10 corresponds to the 10% of points with the highest usage predictions, while 1 corresponds to the 10% of points with the lowest usage predictions. This shows that low-power points generally are more predictable than high-power points, probably because as Figure 3’s summary statistics show, many datapoints contain refrigeration and/or HVAC usage. This difference is minor though, and in any case 78% of real power is concentrated in the top 5 deciles: deciles which show very similar average error magnitudes. Therefore, this framework will be suitable for per-home analysis. It may be expected, though, that higher values will be underestimated by a lognormal model given these error results: since the mean of a lognormal distribution is e^(μ + σ2/2), points with high-magnitude error terms under a lognormal model will have higher average real values than points that were given equal predictions with more certainty.

This nonlinear framework, however, cannot be applied to per-home, per-usage category modeling. Given that many power usage categories (laundry, cooking, etc.) are almost entirely dominated by 0s, each home has very few points within the 8-month training window corresponding to nonzero usage. This means that resulting models of exclusively nonzero power levels will be either overfitted or overdetermined due to the high ratio of degrees of freedom to points surveyed. Given these challenges in modeling the distributions of these per-category signals, linear models were used for this granularity. By contrast, when groups are analyzed, these analytics challenges are sidestepped because much of the skew in power distributions disappears, as Figure 6 shows. After all, by virtue of the central limit theorem, summing or averaging many independent non-normal signals will yield a progressively more normal signal (Granville, 2021). So, for per-group analysis, linear regression of target variables is suitable. Overall, the data flowed through R according to Figure 7’s activity diagram: at the four chosen granularities (total usage by home, total usage by group, and per-category usage by group), models were created at the granularities of the datasets they rely on. Then, these datasets were aggregated to generate average power levels—both real and predicted—for each timestamp. And finally, six regression models were created predicting average aggregate power usage as a function of the average of the low-level models’ predictions.

The topline results of these tests— test R2, validation R2, and validation RMSE—are shown Figure 8. Figure 9 shows the correlation matrix between the regression models’ predictions, while Figures 10-15 show each of the aggregate models alongside their respective real vs. fitted point scatterplots. Parenthetically, the R2 printouts from R are incorrect, so those values were calculated separately as the square of correlation (Josliber, 2015). Firstly, it’s notable that the accuracy values are so similar across all six regressions: both breaking down power usage by category and showing home-level data did not yield additional predictive value. The only model which showed consistent improvement vs. the benchmark per-group/total usage-based regression was the per-group/model-of-categories regression, and even then, the improvement was negligible (an improvement of R2 of ~0.003). This is highly predictable given Figure 9’s correlation matrix: except for the per-home/total usage-based model—which uniquely used nonlinear regressions—all models’ aggregate predictions correlate with each other with coefficients >0.996. Even the scatterplots in Figures 11-15 corresponding to these five linear models are virtually identical. On top of that, both regressions which use per-category predictions as separate regression inputs appear to be overfitted: for both models, 5/8 categories have coefficients so uncertain as to have standard errors ≥ 0.67, where a 1-1 relationship is expected. In short, these coefficients wildly vary from 1 without providing either improvements in predictive value or intuitive reasons for deviating. So, all taken together, it appears that:

1. All models yielded highly similar accuracies to the benchmark model (per-group modeling of total usage data without considering usage categories).
2. Linear models at all tested granularities yield essentially identical outputs.

These takeaways imply that disaggregated data is not useful for aggregate power usage modeling.

## 8.2 Neural Network Analysis in Python

Using the attached neural\_net\_power\_usage\_analysis.py code, neural network analysis was performed. Models built in this manner were fed into the activity diagram shown in Figure 16: much like in the regression analysis section, neural network modeling was done at a low level of analysis and then aggregated by timestamp. However, unlike in the regression section, models that regressed the per-category models together were not created due to the poor results and lack of intuitive justification of that transform. Figures 17 shows the model results, while Figure 18 shows the correlation matrix between models. Figures 19-23 show the top-level linear models and aggregate scatterplots for each of the four pathways specified in Figure 16’s activity diagram. Firstly, it is noteworthy that accuracy is much higher for these neural networks than for regressions: R2 values of ~0.7 were achieved via neural networks while regressions hovered around 0.4. This is likely due to the nonlinearity of relationships between weather and power usage: neural networks have sufficient degrees of freedom to identify those nonlinear relationships on a sufficiently large training set. However, with the per-home models that specificity appears to have come at the cost of robustness: model quality degraded once homes were modeled rather than clusters. This is unexpected, but understandable: each home is highly volatile and says little about any other home in particular; zooming in made it harder for the neural network to find the gradient function. This is confirmed by Figure 18’s correlation matrix: while all models have similar outputs (correlations > 0.92), the minimum correlation is between the home\_total and home\_category models: they have more in common with the group-wise models with each other. This implies that there is a lot of noise and training-dependent volatility: the group-wise models are more accurate, predictable, and robust.

# 9.0 Conclusion

Given these results, high-granularity power usage data—whether at the per-home or the per-appliance level—is simply not a basis for improved power models vs. aggregated data, at least not if one uses regressions as one’s basis for analysis. Essentially, by locally aggregating power usage data—that is, by summing the power signals of dozens of homes with similar locations and then predicting usage within those collectives—one can build power usage models that are more robust, less computationally expensive, and at least as predictively useful as models which attempt to closely model each residence as an individual agent. And while ideally aggregation should be performed between houses with similar prior power signals to retain as clear a collective power signal as possible—thus meriting an unsupervised clustering model for utility customers—aggregating random houses within each county was sufficient to retain near-identical model outputs. So, if the aggregation is between houses that are close enough to experience the same local weather conditions, in practice the useful portions of per-home patterns are preserved post-aggregation, at least when one is predicting usage across a wide service area rather than predicting the usage levels for specific users.

# 10.0 Recommendations

To be sure, per-home data has many modeling uses: for example, anomaly/fraud detection is an important application where data must be available at as high a granularity as possible to both preserve and isolate the signal (Banik et. al., 2023). But it appears to not be applicable for aggregate modeling: using power data at that granularity brings both PII security risks and big data modeling challenges without countervailing benefits. Given these findings, data minimization procedures should be followed in this domain: for power usage modeling analytics pipelines, the power signal should be locally aggregated as early in the pipeline as possible (Ancell, 2022). This does not preclude the existence of high-granularity pipelines existing elsewhere, but it does minimize the organization’s attack space to the areas/pipelines that are justifiable by model improvements, compliance, or other organizational needs (Ancell, 2022). As things stand, applying a feed-forward neural network to a locally clustered power dataset appears to be the optimal approach for aggregate power modeling.

# 11.0 References

Ancell, B. (April 26, 2022). 5 Key Principles for Data Security. <https://www.whymeridian.com/blog/5-Key-Principles-for-Data-Security>

Banik, S., Saha, S. K., Banik, T., Hossain, S. M. M. (June 4, 2023). *Anomaly Detection Techniques in Smart Grid Systems: A Review* [Conference Paper]. 2023 IEEE World AI IoT Congress.

Boeckl, K., Fagan, M., Fisher, W., Lefkovitz, N., Megas, K. N., Nadeau, E., O’Rourke, D. G., Piccarreta, B., & Scarfone, K. (June 2019). *Considerations for Managing Internet of Things (IoT) Cybersecurity and Privacy Risks*. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8228

Buchmann, E., Bohm, K., Burghardt, T., & Kessler, S. (April 2013). Re-identification of Smart Meter Data. *Personal and Ubiquitous Computing, 17*(4).

Buonanno, A., Caliano, M., Pontecorvo, A., Sforza, G., Valenti, M., & Graditi, G. (2022). Global vs. Local Models for Short-Term Electricity Demand Prediction in a Residential/Lodging Scenario. *Energies, 15*(6). <https://doi.org/10.3390/en15062037>

DIY Maps. (n. d.). *Free Color-Coded Maps of North Carolina!!* diymaps.net/nc.htm

Granville, V. (July 16, 2021). *Central Limit Theorem for Independent Random Variables*. Data Science Central. https://www.datasciencecentral.com/central-limit-theorem-for-non-independent-random-variables/

GDPR Summary. (December 13, 2018). *Data Minimisation*. <https://www.gdprsummary.com/gdpr-definitions/data-minimisation>

Josliber. (July 19, 2015). *Incorrect R-squared calculated using linear regression in R [duplicate]* StackOverflow. https://stackoverflow.com/questions/31506283/incorrect-r-squared-calculated-using-linear-regression-in-r

Hosking, J. R. M., Natarajan, R., Ghosh, S., Subramanian, S., & Zhang, X. (2013). Short-Term Forecasting of the Daily Load Curve for Residential Electricity Usage in the Smart Grid. Applied Stochastic Models in Business & Industry, 29(6), 604-620. <https://doi.org/10.1002/asmb.1987>

Li, Z., & Fotheringham, A. S. (2020). Computational Improvements to Multi-Scale Geographically Weighted Regression. *International Journal of Geographic Information Science, 34*(7), 1378-1397. <https://doi.org/10.1080/13658816.2020.1720692>

MathWorks. (n. d.). *lognstat—Lognormal Mean and Variance*. https://www.mathworks.com/help/stats/lognstat.html

National Centers for Environmental Information. (n. d.). *U.S. Local Climatological Data (LCD)*. National Oceanic and Atmospheric Administration. <https://www.ncei.noaa.gov/access/search/data-search/local-climatological-data>

National Oceanic and Atmospheric Administration. (n. d.). *Local Climatological Data (LCD) Dataset Documentation*. <https://www.ncei.noaa.gov/pub/data/cdo/documentation/LCD_documentation.pdf>

Natividad, L. E., & Benalcazar, P. (January 2023). Hybrid Renewable Energy Systems for Sustainable Rural Development: Perspectives and Challenges in Energy Systems Modeling. *Energies, 16*(3). https://doi.org/10.3390/en16031328

Neo, H. Y. R., Wong, N. H., Ignatius, M., Yuan, C., Xu, Y., & Cao, K. (2023). Spatial Analysis of Public Residential Housing’s Electricity Consumption in Relation to Urban Landscape and Building Characteristics A Case Study in Singapore. Energy & Environment, 34(2). <https://doi.org/10.1177/0958305X211056031>

Li, L., & Zha, H. (January 2018). Energy Usage Behavior Modeling in Energy Disaggregation via Hawkes Processes. ACM Transactions on Intelligent Systems and Technology, 9(3). <https://doi.org/10.1145/3108413>

O'Leary, Z. (2021). The Essential Guide to Doing your Research Project. Sage Publications.

Mamun, A. A., Sohel, M., Mohammad, N., Sunny, M. S. H., Dipta, D. R., & Hossein, E. (July 29, 2020). A Comprehensive Review of the Load Forecasting Techniques Using Single and Hybrid Predictive Models. *IEEE Access, 8*. https://doi.org/10.1109/ACCESS.2020.3010702

Thorne, S., Mortveit, H., & Marathe, M. (December 14, 2022). *Household-Level Disaggregated Hourly Synthetic Energy Use Profiles for the United States*. University of Virginia Dataverse. <https://net.science/files/a590ac85-c987-4643-84c5-3737891bf136/>

Zhu, J., Liao, Q., Lin, Y., Lei, W., & Gao, J. (2019). A High-Resolution Model for Simulation of Flexible Residential Electricity Demand [Conference Paper]. 7th International Conference on Information Technology: IoT and Smart City, Shanghai, China. <https://doi.org/10.1145/3377170.3377229>

# 12.0 Tables and Figures

**Figure 1**

*Literature Graph*

A screenshot of a computer

Description automatically generated

**Figure 2**

*Weather Stations’ Jurisdictions*

A map of the state of north carolina

Description automatically generated

**Table 1**

*Consolidated Data Dictionary*

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Data Type | Field Size | Description |
| home\_id | int | 15 | 15-digit unique identifier |
| fips\_code | str | 5 | County FIPS code |
| timestamp | str | 22 | Local timestamp of observation |
| total\_kw | float | 5 | Total power consumption |
| hvac\_kw | float | 5 | HVAC power consumption |
| hoth2o\_kw | float | 5 | Hot water power consumption |
| refr\_kw | float | 5 | Refrigerator power consumption |
| light\_kw | float | 5 | Lighting power consumption |
| misc\_kw | float | 5 | Miscellaneous power consumption |
| dw\_kw | float | 5 | Dishwasher power consumption |
| laundry\_kw | float | 5 | Laundry power consumption |
| cook\_kw | float | 5 | Cooking power consumption |
| dry\_bulb\_temperature | float | 3 | Dry bulb temperature (F) |
| wet\_bulb\_temperature | float | 3 | Wet bulb percentage (F) |
| humidity | float | 2 | Relative humidity (%) |
| precipitation | float | 3 | Rainfall (in/h) |
| air\_pressure | float | 4 | Air pressure (in Hg) |
| wind\_speed | float | 5 | Wind speed (mph) |
| oktas | int | 1 | Sky cover (oktas) |

**Figure 3**

*Problem with Using One Large Regression Model*

****

**Figure 4**

*Power Data Summary Statistics, Per-Home Data*

A screenshot of a computer

Description automatically generated

**Figure 5**

*Residuals for Log(Power) Model*

A screenshot of a computer

Description automatically generated

**Figure 6**

*Summary Statistics, Per-Group Data*

A number of numbers on a white background

Description automatically generated

**Figure 7**

*Regression Models Activity Diagram*

**A diagram of a diagram

Description automatically generated**

**Figure 8**

*Regression Model Results*

**A number of numbers on a white background

Description automatically generated**

**Figure 9**

*Regression Models’ Fitted Values Correlation Matrix*

**A screenshot of a computer

Description automatically generated**

**Figure 10**

*Usage Regression Model, Per-Home Data, Total Usage Prediction-Based*

A screenshot of a computer

Description automatically generated

**Figure 11**

*Usage Regression Model, Per-Home Data, Sum of Category Predictions-Based*

*A screenshot of a computer

Description automatically generated*

**Figure 12**

*Usage Regression Model, Per-Home Data, Regression of Category Predictions-Based*

*A screenshot of a computer

Description automatically generated*

**Figure 13**

*Usage Regression Model, Per-Group Data, Total Usage Prediction-Based*

*A screenshot of a computer

Description automatically generated*

**Figure 14**

*Usage Regression Model, Per-Group Data, Sum of Category Predictions-Based*

*A screenshot of a computer

Description automatically generated*

**Figure 15**

*Usage Regression Model, Per-Group Data, Regression of Category Predictions-Based*

A screenshot of a computer

Description automatically generated

**Figure 16**

*Neural Network Models Activity Diagram*

*A screenshot of a computer

Description automatically generated*

**Figure 17**

*Neural Network Model Results*

*A screenshot of a computer

Description automatically generated*

**Figure 18**

*Regression Models’ Fitted Values Correlation Matrix*

*A screenshot of a computer

Description automatically generated*

**Figure 19**

*Aggregate Point Scatterplots for Neural Network Models*

***A screenshot of a computer screen

Description automatically generated***

**Figure 20**

*Topline Regression, Per-Group, Total Usage-Based*

***A screenshot of a computer

Description automatically generated***

**Figure 21**

*Topline Regression, Per-Group, Sum of Category Usage-Based*

***A screenshot of a computer

Description automatically generated***

**Figure 22**

*Topline Regression, Per-Home, Total Usage-Based*

***A screenshot of a computer

Description automatically generated***

**Figure 23**

*Topline Regression, Per-Home, Sum of Category Usage-Based*

***A screenshot of a computer

Description automatically generated***