Final Project

North Carolina Electric Membership Corporation – Load Forecasting

**The Challenge** (per <https://www.sas.com/en_us/customers/north-carolina-electric-membership-corp.html>)**:**

*Forecasting power demands and consumption patterns on an hourly, daily, and monthly basis for up to 30-40 years to ensure smart capital and construction decision and the reliable availability of lower-cost power for more than 950,000 North Carolina households.*

**Overview:**

The North Carolina Electric Membership Corp. supplies more than 950,000 households in 93 of the 100 counties within the state of North Carolina. Rapid growth within the state is requiring more detailed and quicker load forecasting to make decisions on how, when, and where additional generation will be needed to keep up with the population growth.

As population growth increases throughout the state, there will be an increase in load demand. With an increase in load demand, the current generation portfolio needs to be optimized to assess if it can keep up with the increased load demand. If the current generation portfolio cannot meet the increased load demand, new generation assets must be built and brought online to supply the increased load demand. This type of analysis is not uncharted territory for a utility. In the past, doing this type of analysis would be extremely costly in time and dollars. There is a large variety of data that needs to be collected and used as inputs to the different models. Previously, most of the calculations would be been done by hand or rudimentary software packages. Given the new tools and technologies available in the present day, this type of analysis can be performed quickly, more accurately, and will have a significant cost savings aspect.

The three criteria that need to be analyzed are:

* Population growth
* Potential increase in load demand (customer usage)
* Potential increase in generation

In analyzing the above criteria, the following factors also need to be considered for each criterion:

* Data Used
* Data Collected
* How often to refresh the data and re-run the model

**Abstract:**

The starting point of this analysis is to create a model that predicts population growth across the 93 counties that are provided power. Once there is a population growth model, the results from that model will need to be classified as different population types and assigned a percentage of total growth. Next, a customer usage model will need to be created based on data from each population type and will need to include the effects of seasonality. This will determine how load demand increases with population growth. Finally, the current generation portfolio will need to be optimized to see if it can keep up with the predicted load demand forecast for the different population types. If the current generation portfolio cannot keep up with the forecasted load demand, a model will need to be created to predict additional, cost effective power generation.

It is assumed that the data needed for each model is cleansed, and scaled appropriately prior to being used in the model. It is also assumed that all data is split into training, testing, and validating data sets as necessary to ensure the validity of the models used in the final analysis.

**Population Growth Model**

The makeup of North Carolina is a mixture of large cities, numerous suburbs, vast amounts of farmland, mountains, and beach. Given that the NCEMS serves 93 of 100 counties within North Carolina, it can be assumed that the NCEMS serves a variety of customers in different living situations. It is also a fair assumption to state that the population growth is not equal between the various living situations. To address the issue of non-equal growth, the living situations need to be clustered so that a more accurate population model can be predicted.

The data needed for the clustered model could get gathered from state and national public records, including but not limited to marriage/divorce paperwork, real estate deeds, and birth certificates. The type of data needed would include:

* Marital status
* Number of children
* Address (including county)
* Median Family Income (by country)
* Living arrangement (own/rent)

The data necessary for the population model has the potential to be sparsely populated due to the human input needed to collect the correct data. Due to potential sparsely populated data, a means for imputation should be considered. One suggestion would be to use that combinations of the data that is available to impute for the missing data with new categorical factors. With the above data split into training, testing, and validating data sets and using k-means clustering, the living situation of the population could be ideally clustered into the following categories:

* Urban
* Suburban
* Rural

It is possible that more than three clusters are distinguished and that should be taken into consideration and properly modeled. After proper clustering, a regression model would be built to determine if the population can be accurately modeled based on clustering. Once a proper cluster and regression algorithm is defined, a population growth model could be generated for each category. A first try at modeling population growth is to state that population growth is exponential. A combination of population data by county over the previous 30 years and the output of the regression model could be used to determine the total population of each cluster category (by knowing which counties are represented in each cluster category). Using the population data by county over the previous 30 years, a population growth rate could be calculated using:

Where:

* P – final population
* Po – initial population
* r – growth rate
* t – time between initial and final population

The calculated growth rate could be used to predict the population growth by country for the next 30 years. The population growth would then be distributed back to the cluster categories to determine the predicted 30-year population for each category, by value and by percentage.

This model should probably be re-run every 10 years. This would coincide with the US Census and could account for things that could cause a mass change in population for the positive or the negative (new, large businesses coming to town, discovery of contaminated water source).

**Customer Usage Forecasting**

A starting point for the load demand forecasting would be to create a CUSUM model to see where power usage started to increase over the last 30 years. The CUSUM model would need to be run for each population category and include a calculation of the load factor for each population category. Load factor is defined as:

The threshold value for the CUSUM model would be the mean load factor for each population category. The output of the CUSUM model could then be compared to the population growth of each population category from the last 30 years to analyze if there was an increasing trend of power usage and population growth. This would confirm the hypothesis that there is a trend of increasing power usage with an increase in population, and allow a value to be assigned to that trend.

The data needed for the load demand forecasting CUSUM model is all data that will be in the power company databases. Due to the large amount of data generated at a power company, it will probably be difficult to find 30 years of daily data. At some point, historical data is probably rolled into a less granular time frame for storage capabilities. This less granular data should be taken into consideration when running the various models. Depending on the model being run, the data may all need to be on the same time scale for an appropriate output. It is assumed less imputation would be needed on the power company’s data, but if imputation is needed, means and medians would be a good starting point for filling in missing data. The data needed for the CUSUM model, includes:

* Previous customer power usage (monthly by county)
* Peak load (largest load for each population category)
* Time (in terms of billing cycle days)

The CUSUM model should be run over the months for the previous 30 years for each population category.

A regression model to predict power usage should be built with the following factors:

* Previous customer power usage (ideally daily, but could be monthly by county)
* Temperature (daily average for the county if daily power data is available, else monthly)
* Population (yearly average for the county)
* Day of the week (binary, 0 – weekday, 1 – weekend)

An exponential smoothing model with trend and forecasting should be created to account for peak power usage per day. This type of time series modeling is good for short term forecasting, which would allow the estimated daily peak usage to be calculated. The exponential smoothing model with trend and forecasting would need to be optimized with the training data set to find the best values of α, β, and γ for the model. The data needed for this time series model includes:

* Peak usage (daily if available, but could be monthly by county)
* Peak usage time (hour of the day)
* Temperature (daily max for the county)

The exponential smoothing model with trend and seasonality would need to be re-run for each month to estimate the daily peak usage. The seasonality effects must be broken down by counties because two counties that may fall into the same category may be at temperature extremes (a rural mountain county and a rural coastal county).

The output of the peak usage time series forecasting model would be combined with long term regression model to create a complete estimated power usage model including seasonality and accounting for the daily habits and routines of the customers. Once the completed model is generated, the predicted county population, forecasted temperatures, forecasted daily peaks, and future days would be inputs to predict future customer usage.

A new consideration that may want to be added to the power usage model is the idea that while the population may not be growing at a rapid pace, the consumer electronic industry is growing at a very rapid pace. In previous models of power usage, this phenomenon did not have to be accounted for as there were not a lot of personal electronic devices in the past. The current US average for personal electronic devices per person is around four. It should be expected that this number will increase in the future, and could possibly affect future power usage estimations. This consideration could be modeled using a Neural Network/Deep Learning. This algorithm would allow a model to be built that is learns to account for the multiple devices of everyone within a household, which sums to the power demanded by a residence.

**Power Generation Model**

As the population grows and load demand increases, there may be a need to build new sources of generation to keep up with the growth. An optimization program could be built to determine the maximum demand factor that still ensures high reliability and high availability. Demand factor is defined as:

For the optimization model, maximizing the demand factor would be the objective function. Based on the optimization, a demand factor threshold could be set. The threshold for the demand factor should be set conservatively as the cost of using an aggressive value is loss of power to customer, which negatively affects the bottom dollar of the company. If the solution to the optimization problem determines that long term demand was going to exceed the threshold, plans for building new generation should be considered. This optimization program would need to be run on all the sources of generation and have the aggregate of the optimization results evaluated to determine the need to build new generation. This model should be re-run every time the customer usage forecasting models are re-run. This would ensure that the generation would always keep up with the customer usage forecasting.

Like in the customer usage model, the data needed for the power generation model would be available from the databases of the power company with the same assumptions. The data needed to compile the optimization model includes:

* Max generation (daily)
* Average generation (daily)
* Availability (daily)
* Reliability (daily)
* Capacity factor (daily)
* Average temperature (daily)
* Precipitation (daily)
* Wind speed (ideally hourly, but could be daily)
* Lake/Reservoir level (daily)
* Number of generators online (daily)
* Average number of generating hours (daily)
* Type of generator (coal, fossil, solar, wind, nuclear, hydro)

Some constraints that should be considered for this model include:

* Minimum values that must be met for daily availability and reliability
* Maximum value for capacity factor
* Solar and potentially wind power not being viable sources of generation during inclement weather (overcast, precipitation)
* Solar generation not being available in the evening
* Hydro generation only being available when the source level is greater than the minimum required level for operation

Once it is determined that a new generating source is needed, the location for the new generating source will need to be chosen. When selecting the type of generator to build, data from the optimization problem should be used to aid in determining the correct type of generator to build for the specific location chosen. It could also be important for the company to reach out the customers through a survey to determine if there are any types of generators people are opposed to, and if so, are they potentially willing to pay more for their power to avoid these types of generators. The idea behind this survey is that it may be cheaper to build a coal fired plant to produce the necessary additional power, but people may be against the byproducts and would be willing to pay more for a natural gas plant. This survey could be performed using a series of A/B testing to gather customer responses.

**Conclusion:**

Creating these models is a cost benefit to the NCEMC as it will allow them to quickly assess growth and whether they are currently capable of keeping up with the growth while optimizing the usage of the current assets. Once the models are created, they do not require a lot of manpower to maintain and run repeatedly, which is also a cost benefit to the NCEMC. This challenge involves collecting and compiling a variety of data to use in numerous models. This example is a very high-level attempt at solving the problem with analytics. Some additional details that could also be considered in this problem is the effect on transmission and distribution systems. There is a potential that these systems should be analyzed to ensure they could accept the potential additional generation to accommodate the load demand growth. The outputs of these models should be used in conjunction with other economic and environmental models to truly understand how to provide the customers with reliable power for the future at a reasonable cost.