NOVEC
Project Proposal

Authors:
Abdulrahaman Alsabti
Kevin Mitchell
Dan Package

10/02/2014

Contents

I. Introduction

a. Background

Northern Virginia Electric Cooperative (NOVEC) is a power wholesaler headquartered in Manassas, Virginia. NOVEC gives power to in excess of 150,000 clients over different regions. NOVEC is obliged to provide electrical power for any level of client demand. With that requirement as a boundary condition, NOVEC buys power from retail power suppliers. Temperature changes, primarily amid the late spring months, are a huge driver for expanded power request. In order to minimize the amount of purchases without overcompensating with excessive bulk purchases, NOVEC has developed a forecasting model that estimates future energy purchases. NOVEC leverages forecast model insights to inform the magnitude (kilowatt-hours or kWh) and length of energy purchases from suppliers. Economic metrics included in the model seek to characterize the basic load by capturing economic growth or decline in the Northern Virginia area. The basic load is the energy requirement based solely on the size and typical consumption of customers, the number of which changes with time. To reasonably determine the size and growth rate of customers, local weather data is collected and used to remove the effects of weather on historical energy purchases. NOVEC prepares monthly long term (by month for 30 years) but they only look at the first 10-20 years. They also provide daily short term forecasts (for a period of three days). The extensive forecast period is important to accommodate the complex capital planning, development, infrastructure efforts required to expand electric power capacity, and delivery in accordance with federal and state regulations.

NOVEC's models consider customer characteristics in addition to weather-related and economic factors. The models currently include the following categorization of customers:

- High volume customers (aka "needle-movers", ~5-10)
- Non-high volume residential customers (~140,000)
- Non-high volume non-residential customers (~12,000)

Historical monthly usage data (the basis of power bills) for each residential customer drive a baseline load for a residential customer. The models then represent each non-residential customer as a multiple of a residential customer. NOVEC assigns each customer a billing rate code that is used to drive bill rates based on the expected consumption pattern of the customer. The models do not consider these billing rate codes in the forecast calculations.

In 2011, NOVEC started a program to replace more than 1,100 regular meters with survey meters. A survey meter captures hourly usage in kWh. NOVEC has supplied this survey meter data to the project team along with the map coordinates and billing rate code associated with each meter.

b. Problem Statement

Currently NOVEC purchases power from power suppliers depending on the customer demand. Usually there is an amount of power that has been purchased but hasn't been consumed by any of NOVEC's customers. This causes an increase cost of providing power to customers, because they are ordering more power than needed. NOVEC is developing a system that can better forecast customer usage. However, they understand that categorizing customers in more than two categories, residential and non-residential, can increase the accuracy of their load forecast. NOVEC is requesting to provide a better algorithm for customer categorization that can help in better forecasting in the future.

NOVEC is also interested in learning if the survey meter population is representative of their overall customer population.

c. Stakeholders

The primary stakeholder in this project is NOVEC who is the project sponsor. The primary points of contact at NOVEC are Bryan Barfield and Bob Bisson, who will be working with the project team and providing the required information for them. As representatives of the NOVEC customer base, Mr. Bisson and Mr. Barfield, expressed the need for better load purchasing forecasts in order to minimize cost to the consumer. As the project advisor, Prof. Hoffman is a stakeholder who wants the sponsor to be satisfied with the team's success.

d. Scope

The scope of the project is two-fold. First, the team will devise an algorithm that can categorize NOVEC's current and prospective customers into multiple categories. The team will consider billing rate code and the survey meter data in devising this algorithm. The use of these categories will help NOVEC's current forecasting system to better forecast load, which will lead to decreasing the cost of energy purchased from other power suppliers. We intend to categorize the customers via data regression and clustering. Data regression and clustering discovered around the major trend analysis will be used by NOVEC personnel after the completion of this project to feed additional independent variables in their current forecasting models. If there is additional time remaining in the project, the team will investigate ways to correlate geographic grid data with public domain information to provide further approaches to categorize customers based on their location.

Second, the team will assess the extent to which the survey meter sample represents the overall population. If the team determines the sample does not accurately represent the overall population, the team will provide recommendations on how to structure future samples.

II. Requirements

1.0 Deliverable Requirements

1.1 The Team shall develop customer categorizations for NOVEC.
1.2 The Team shall provide status briefings to the course instructor.
1.3 The Team shall have meetings with the sponsor when needed.
1.4 The Team shall provide a final report of this project to the course instructor.
1.5 The Team shall provide a final report of this project to the sponsor.
1.6 The Team shall present the results of this project to the GMU SEOR faculty, students and sponsors on December 12, 2014.
1.7 The Team shall design a website containing all final deliverables.

2.0 Project Requirements

2.1 The team shall choose technologies for data analysis that minimize cost while maximizing ability to discover the patterns expected to emerge from the sample data.
2.2 The team shall develop a strategy to categorize the various customers.
2.3 The categorizations shall provide NOVEC an additional independent variable to feed into the current forecasting system.
2.4 The categorizations from the sample data provided shall reflect the overall population of the customers.
2.5 The team shall provide all data transformation available to NOVEC for future use.
2.6 The team shall recognize "like" customers in terms of their usage or any other factor.
2.7 The team shall use load shape to compare and contrast different customers

III. Approach

The team expects the analysis process to involve an iterative approach using data clustering algorithms and evaluation of clusters treated as customer categories to scale the sample population up to the load across the total customer base. We will first explore billing rate code as a categorization/clustering variable. If billing rate code proves to be a poor classifier for customer load, the team and NOVEC will identify other available data that may provide better results within the time and resource constraints of the project. The geographic grid data already provided by NOVEC is an example of such additional data.

In parallel to the data clustering for customer categorization, the team will assess the ability of the sample (i.e. survey meter data) to make conclusions on the overall population.

The team plans on using the new categories to enhance NOVEC's forecasting system. We will be able to test the new categories by using the old historical data against the newer historical data.

For either activity, the team will retain and package data processing and statistical scripts and code for delivery to NOVEC going forward.

a. Data Processing Automation

Given that NOVEC expects to enjoy continued growth in its customer base in the coming years, this project should result in an easily reusable method to process the raw data generated by the survey meters. While the revised customer categories should improve forecasting accuracy in the near-term, the team expects that this approach should be employed at regular intervals in the future to maintain that precision. Initial prototypes for aligning data over customer ID and time of collection have involved Java command-line executable code, but whatever form the data pre-processing methods take, the team will provide documentation how to execute, modify and maintain the code and expected file/folder structures. The general approach has and will most likely involve taking the vertically-aligned data with repeating customer identification and dates within the same file, to a more tabular listing of unique dates and customer identification in row-and-column format further binned and separated by calendar month and by billing rate code.

The team anticipates using R to perform statistical calculations, processing, and data visualizations. Like the Java data processing programs mentioned above, the team will provide documentation and R code to NOVEC for use going forward.

b. Data Regression and Clustering

Several software packages exist that perform different data clustering analysis methods such as K-means and Expectation-Maximization (EM) algorithms. The team will work with John Miller, a George Mason faculty member, who will serve as an advisor conversant in statistical analysis of data to choose an approach that bears the most promise for extracting new customer categories. Professor Miller will assist with determining methods that will reduce the signal to noise ratio in the data such as the possibility of de-normalizing the data through subtracting out minimum threshold wattage common to all customers, establishing different mean usage levels, etc. In general, the team feels that the 24-hour load shape adjusted for day of week variance holds the most potential for success. The software tools that feature these types of analysis, such as R and Weka, have open source versions at no cost to the user, and the team will evaluate these alternatives with the faculty advisor.

Once the usage data has been organized into meaningfully correlated data grouped by customer type and calendar month as described in the previous section, the team will proceed with a two-phase approach to evaluating the data. Phase one will consist of establishing a coefficient of variation across data points within each existing billing rate code for a fixed month of the year (possibly holding other variables constant such as day of week or other factors that influence usage). After we have calculated the coefficient of variation for the billing rate codes, we will

analyze it noticing if they have little or high variance. Whether the coefficient of variation is high or low, it will influence the path of phase two. If the coefficient of variation is low, the team will investigate similarities between existing billing rate codes and the possibility of merging one or more rate codes into a shared category based on usage behavior. If the coefficient of variations high, phase two will involve methods to determine new categories unrelated to rate code and require a greater degree of input from Professor Miller as to how to select new hypothetical means/averages around which to cluster customer data. The final result of which will either be increased confidence in using the existing customer billing rate codes as categories to feed short term forecasting, or some number of fewer or more customer categories suggested by the team.

C. Test and Evaluation

Discussions with NOVEC personnel have called attention to the need for a quantitative measure of success for this analysis effort. Their suggestion was to simultaneously evaluate a) the categories developed through the clustering analysis along with b) the validity of the sample size and representation of the different customer types within the sample as compared to the true population of customers. If the total load registered by the survey meters multiplied by the ratio of total population to survey size does not match total load across the system, the error could be corrected by calculating a weighted scaling factor for each of the new categories and estimating the ratio of newly discovered categories within the total population (likely different from the percentage of each category within the sample population). As an example: if the total load across 800 customers is “X” and the categories break down such that Customer type A is 15% of 800, Customer type B is 35% of 800 and Customer type C is 50% of 800, then scaling the total load “X” up to the full 150,000 customer population while assuming the same breakdown of categories within the total population will likely not match the true total load as measured with non-survey meters. The team should work with NOVEC personnel to identify a likely correlation between the load shape categories and a possible type of customer to estimate the scaling factor illustrated in the table below. It follows that if customer categorization will help refine power purchasing forecasting, then that prediction is most sensitive to the true number of customers in each category across the total population versus the number in each category found in the sample size as indicated by the column highlighted in red. One method to identify this is to investigate identifying attributes about the customer type based on public domain data as described in the next section.

Sample	800	
Total Pop	150,000	Total : Sample
		Ratio = 187.5

Category	% cust in sample	# cust in sample	Total load	Sample Scaling Fact	# cust in total pop	Total Load
A	15%	120	20000	1	22500	3,750,000
B	35%	280	10000	1	52500	1,875,000
C	50%	400	7000	1	75000	1,312,500
		Tot Samp	37,000		Tot Scaled	6,937,500
					Load Ratio	187.5

Discussions with Professor Miller have highlighted that the best case scenario is that NOVEC personnel will be able to provide the original methodology for choosing survey meter installation sites and that the scaling will be fairly evident once that is known. However, in the absence of a known methodology for sample selection, certain assumptions can be made that will allow the scaling approach to remain feasible for evaluation of the final categorization of customers with some degree of definable risk. Additionally, depending on the outcome of phase one, the team plans to investigate additional independent variables that might feed customer categorization such as their location within the coverage area.

NOVEC tracks its customers with grid coordinates that were included with the raw data provided to the team. If load data does not scale correctly, the team will investigate the possibility of correlating customers' grid location with residential data (such as home value, projected income or other census data) in order to help adjust the scaling factors described earlier.

IV. Project Plan

a. Work Breakdown Structure

A Work Breakdown Structure (WBS) was developed to assist in planning, evaluating, and managing project tasks. The WBS has been decomposed into five components: project management, deliverables, front end, back end, and solution. Project management consists of planning, team meetings, Earned Value Management (EVM). The purpose of these tasks is to ensure the project team remains focused on sponsor needs, within budget, and on time. Deliverables include presentations, written reports, and a project website. The front end analysis consists of problem definition, scope, and requirements. It also includes data analysis, which is critical for the back end analysis and development that encompasses the design, and testing. The solution includes analysis of results and group recommendations for the problem

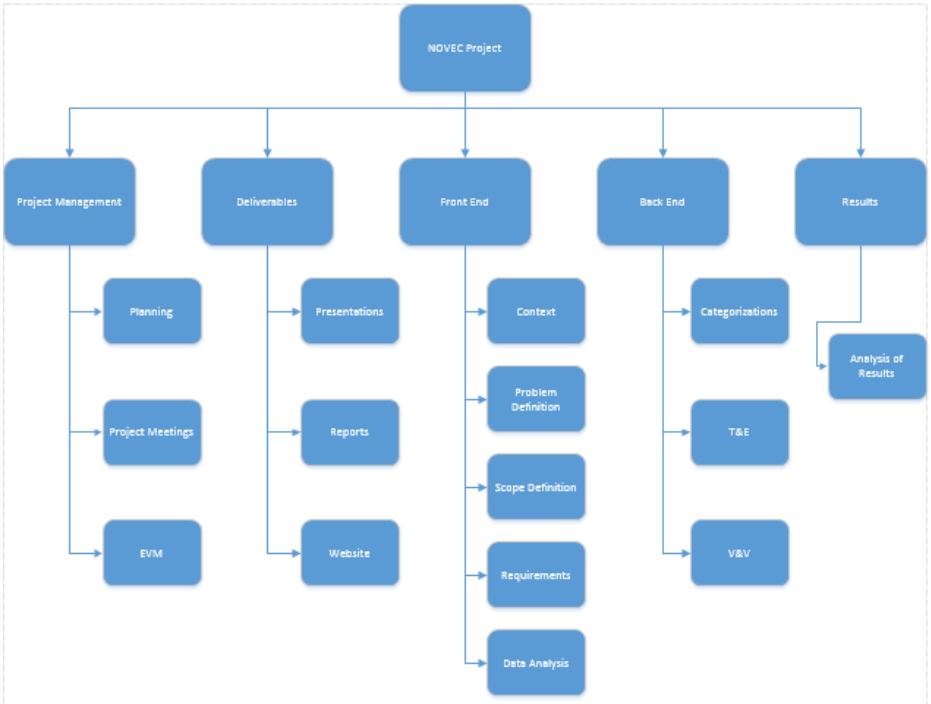
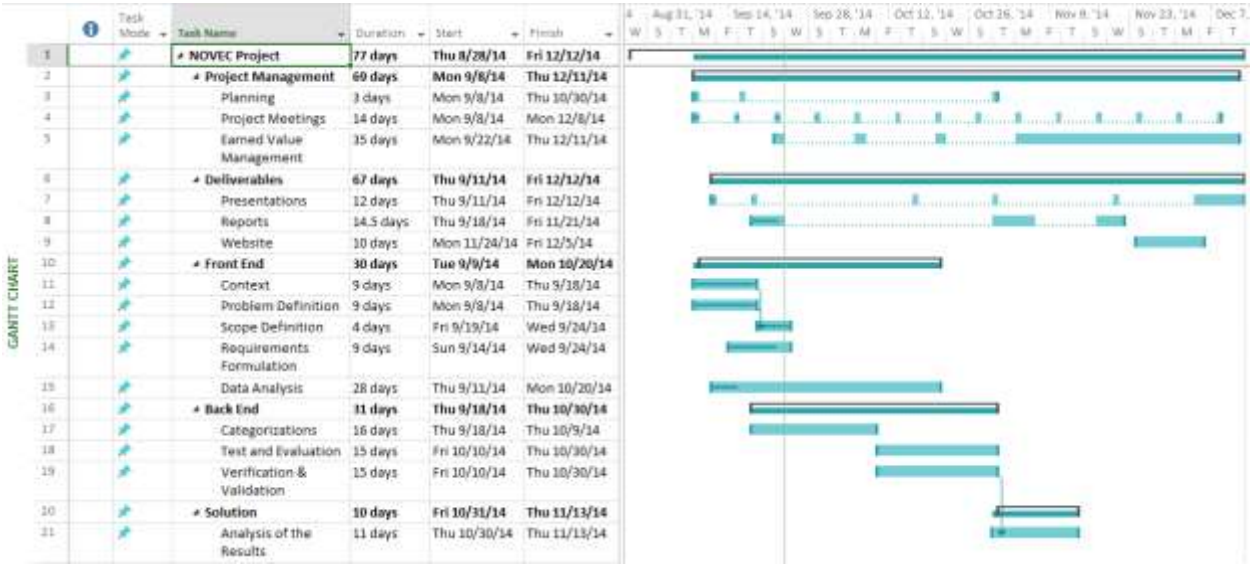


Figure 1. WBS for NOVEC project

b. Schedule

The project schedule was implemented as a Gantt chart. The schedule was set to 16 weeks. The duration of each task in the WBS was estimated. The project progress will be tracked throughout the semester using EVM.



	Task Mode	Task Name	Duration	Start	Finish
1		NOVEC Project	77 days	Thu 8/28/14	Fri 12/12/14
2		Project Management	69 days	Mon 9/8/14	Thu 12/11/14
3		Planning	3 days	Mon 9/8/14	Thu 10/30/14
4		Project Meetings	14 days	Mon 9/8/14	Mon 12/8/14
5		Earned Value Management	35 days	Mon 9/22/14	Thu 12/11/14
6		Deliverables	67 days	Thu 9/11/14	Fri 12/12/14
7		Presentations	12 days	Thu 9/11/14	Fri 12/12/14
8		Reports	14.5 days	Thu 9/18/14	Fri 11/21/14
9		Website	10 days	Mon 11/24/14	Fri 12/5/14
10		Front End	30 days	Tue 9/9/14	Mon 10/20/14
11		Context	9 days	Mon 9/8/14	Thu 9/18/14
12		Problem Definition	9 days	Mon 9/8/14	Thu 9/18/14
13		Scope Definition	4 days	Fri 9/19/14	Wed 9/24/14
14		Requirements Formulation	9 days	Sun 9/14/14	Wed 9/24/14
15		Data Analysis	28 days	Thu 9/11/14	Mon 10/20/14
16		Back End	31 days	Thu 9/18/14	Thu 10/30/14
17		Categorizations	16 days	Thu 9/18/14	Thu 10/9/14
18		Test and Evaluation	15 days	Fri 10/10/14	Thu 10/30/14
19		Verification & Validation	15 days	Fri 10/10/14	Thu 10/30/14
20		Solution	10 days	Fri 10/31/14	Thu 11/13/14
21		Analysis of the Results	11 days	Thu 10/30/14	Thu 11/13/14

Figure 2. Project schedule for NOVEC