Treatment Effect of Repairs to an Electrical Grid

Leveraging a Machine Learned Model of Structure Vulnerability

Rebecca J. Passonneau
Center for Computational Learning Systems
Columbia University
New York, NY, USA
becky@cs.columbia.edu

Cynthia Rudin
MIT Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA, USA
rudin@mit.edu

Axinia Radeva
Center for Computational Learning Systems
Columbia University
New York, NY, USA
axinia@ccls.columbia.edu

Ashish Tomar
Center for Computational Learning Systems
Columbia University
New York, NY, USA
atomar@ccls.columbia.edu

Boyi Xie
Department of Computer Science
Columbia University
New York, NY, USA
bx2109@columbia.edu

ABSTRACT

Consolidated Edison of New York is the utility company that provides electrical power to New York City. As of late 2004, the Public Service Commission of the State of New York requires Con Edison to inspect all electrical structures in the power grid, such as manholes and service boxes, at least once every five years. In our previous work, we implemented a process to assemble a wide array of textual and semi-structured data and from it to produce a machine learned ranking of vulnerability of secondary structures for a given year. For the work reported here, we were asked to analyze the inspections data from the first five-year cycle for Manhattan, to determine if a relation could be found with our structure ranking. This paper describes the results of causal inference using the inspections data. We tested whether repairs carried out in response to inspections have a positive impact on the health of structures. Results indicate a highly significant effect in which repairs triggered by inspections reduce events in the immediately following year. We then partitioned the structures into distinct strata based on their structure rank, and again tested for a treatment effect of repairs. This stratified analysis yields a decreasing incidence of events for each next stratum with lower vulnerability. The results both confirm the empirical adequacy of the structure ranking we produce, and demonstrate its utility for assessment of maintenance and repair activities carried out on the secondary grid. In turn, these results can have a positive impact on sustainability through more directed allocation of resources, by allowing structures that are more vulnerable to be inspected earlier.

SustKDD 2011 August 2011, San Diego, CA, USA
Copyright 2011 ACM 978-1-4503-0840-3 ...$10.00.

1. INTRODUCTION

The term *power grid* typically refers to the transmission network of high voltage power. In the United States, there are three major transmission networks which have a complex system of automated sensors and devices for monitoring and managing power load and flow. Low voltage distribution networks are fed by the transmission grid and primary distribution network, and provide power directly to customers. The network consist of structures, such as manholes and service boxes, that provide direct access to the cables connecting structures to each other and to customers. There are primary cables that feed power to the distribution network (feeders), the secondary main cables that distribute power throughout a region, and the service cables that provide power directly to customers. Figure 1 (from the Wikipedia article on Electric power distribution at `http://en.wikipedia.org/wiki/Electric_power_distribution.`) illustrates that high voltage transmission cables (blue) provide power to distribution substations, where transformers step the voltage down for distribution (green) along primary distribution cables that feed power to the secondary. In the New York City setting we investigate, below-ground feeder cables (not shown here) distribute power to underground transformers that reduce the voltage to secondary levels (typically 120 volts) to provide power to customers. Compared to the transmission grid, the *secondary* networks are not as fully instrumented. As a consequence of this lack of precise information, and of the redundancy in the network designed to prevent service interruptions, when there is a failure from the primary to the secondary, such as a feeder failure, or a failure within the secondary, such as the loss of a secondary main, *there is no obvious indication that there is a malfunction* [15]. The sustainability problem we address is how to leverage existing offline data in novel ways to assess the health of the secondary network in part of New York City, and how to measure the impact of resources allocated towards maintaining that health. This paper reports how we analyzed a dataset consisting of reports documenting inspections of structures in the secondary grid with a

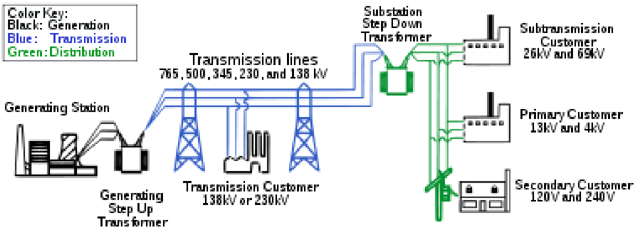

Figure 1: The electrical transmission and distribution system

coombination of techniques for data mining, machine learning, and causal inference from observational data. Our long range goal is to help determine the optimal prioritization for inspecting structures in the secondary grid, which we are addressing in ongoing work. Here we describe results that are a precondition for such optimization: a quantitative assessment of the benefit of inspecting structures, and performing consequent repairs.

As of late 2004, the Public Service Commission (PSC) of the State of New York requires at least once in every five years that Consolidated Edison inspect all structures, including manholes and service boxes, in the electrical networks that provide power to New York City customers. The first five year inspection program was completed in 2009. In our previous work we implemented a process to assemble a wide array of textual and semi-structured data from numerous sources, and from it to produce a machine learned ranking of structure vulnerability for a given year [22]. Each learned model assigns a score and rank order to the thousands of structures in a city region (borough) in order of their vulnerability to serious events. We were asked to analyze the inspections data for Manhattan to determine if a relation could be found with our structure ranking. This paper describes how we applied causal inference to analyze the inspections data, in order to test whether repairs carried out in response to inspections had a positive impact. It represents the first application of our structure ranking work to assessment of the impact of actions, such as repairs, carried out on the grid. We relied on the structure vulnerability ranking to identify distinct *strata* of structures with similar vulnerability. As reported here, we found a statistically significant treatment effect, meaning that repairs have a positive impact on structure health. First, we found that repairs carried out immediately in response to inspections reduce events in the immediately following year. Second, we found a decreasing incidence of events for strata with lower vulnerability, as given by our structure ranking. These results both confirm the empirical adequacy of the structure ranking we produce, and demonstrate its utility for assessment of activities carried out on the secondary grid. This can have a positive impact on sustainability through more directed allocation of resources, by allowing structures that are more vulnerable to be inspected earlier, a problem we address in our ongoing work.

In the next section, we discuss related work on the electrical grid and ranking methods. This is followed by an overview of our project, and our structure ranking model. Then we describe the data on the first five-year cycle of the inspection program mandated by the PSC, our method for performing causal analysis on this observational data, and our results. We conclude with a recapitulation of our results and contribution.

2. RELATED WORK

Our colleagues at the Center for Computational Learning Systems, as far as we know, are the first to use machine learning ranking techniques for applications to the electrical grid [12, 2, 22], though there is much precedent for other maintenance techniques and statistics for use in power engineering (e.g., monitoring the health of power transformers [24, 14]). Our colleagues are concerned with the problem of ranking electrical components in the primary distribution system, specifically electrical feeders, cables and joints, according to their susceptibility to failure. This application differs dramatically from ours in that primary feeders have electrical monitors that provide numerical information in real time; for instance, features are based on electrical load simulations and real-time telemetry data. In contrast, our data consists mostly of historical records written mainly in free-text. It is clear when a feeder fails since an outage of that feeder occurs, whereas it is not always clear when a serious manhole event has occurred. This is a matter of human judgement, and as discussed below, the domain experts do not have complete agreement on serious events.

The problem of feeder susceptibility ranking requires attention especially to the *short-term* timescale before the feeder fails. Thus the problem is an online ranking problem with a short-term crucial timeframe over which warning signs lead to an event. Many other problems dealing with prediction of rare events in continuous time also rely on a shorter time scale, including seizure prediction for epileptic patients [16], or prediction of compiler failures in hard drives (see [17]). In contrast, our task is an offline processing problem that uses a long-term history to predict events that may happen several months later.

There are other works that adopt machine learning techniques for offline prioritization problems, for instance including the prioritization of mutations that cause disease [13], and the prioritization of geographic regions for conservation [4] and species modeling [7]. Those works implement algorithms for density modeling, and prioritization can be done according to the density. In our case, prioritization is the main goal, and thus is addressed as a bipartite ranking problem. Bipartite ranking techniques allow us to characterize the relationships between structures without first estimating the density. There has been a surge of theoretical work on bipartite ranking recently (see, for instance, [10, 1]). One advantage of RankBoost [10] is that there is a theoretical equivalence between RankBoost and AdaBoost [9], its counterpart for classification, meaning that one could obtain a useful classifier from RankBoost [21].

3. SECONDARY MACHINE LEARNING PROJECT

The Consolidated Edison Company of New York City provides electric power through a large transmission and distribution network consisting of a primary grid of high voltage power (transmission) and a secondary network (distribution) that directly serves its 3.3 million customers. The 94,000 miles of primary and secondary cable are connected through 264,000 manholes and service boxes (structures). A structure is a below-ground room that houses cables mounted on racks, with ducts feeding the cables into the structure. Smoldering material resulting from damaged insulation can release gases within the manhole and cause events of smaller or greater magnitude, such as flickering lights in the service area, or an explosion. Our work focuses on the secondary grid, and in particular, on secondary events that affect service reliability (flickering lights) and safety (explosions).

Two key factors pertaining to specification of the time frame we address are that structures evolve relatively slowly over time, and structures degrade relatively slowly over time. Given that a single structure can contain anywhere from just a few secondary cables or none, to hundreds of secondary cables, the large structures will contain cables that have been installed at different points in time. For example, the oldest cable in our databases (described below) was installed in the 1880s, the most recent in the current year. Events involving secondary structures ranging from minor interruptions in service to explosions or fires can result from a slow breakdown in the materials that insulate cables, due to corrosion from weather and salt in the structures. As described in [20], our analyses of events led us to formulate our modeling task on a year-by-year basis.

Beginning in 2006, the Secondary Machine Learning Project has worked with Con Edison to develop methods to model the vulnerability of secondary structures. We have applied these methods to three New York City boroughs[1]: Manhattan (2007, 2010), Brooklyn (2008) and Bronx (2009). Each borough has a distinct network configuration (e.g., radial versus fully connected mesh) and different constituency (e.g., proportion of underground versus overhead structures), thus we model each borough independently. It is a testament to the generality of our approach that the ranking models for different boroughs are distinct (rely on distinct features, and coefficients of these features; see below), thus capturing the distinct properties of individual boroughs, yet have similar performance with respect to blind evaluations. The blind evaluations assess how many structures that experience serious events in a new year are highly ranked [22].

In the work presented here, we tested how rank in our model interacted with the incidence of future events. During the late summer and fall of 2010, we investigated the relation between the Manhattan structure ranking model and data from the first five-year inspections program. Our findings demonstrate the utility of the structure ranking for assessing the impact of inspection-triggered repairs made in 2008 on the incidence of events in the following year. We did not investigate repairs made earlier than 2008 on the recommendation of our collaborators at Con Edison, due to the fact that procedures for carrying out inspections continued to be refined throughout the program, and were insufficiently stable until 2008.

3.1 Raw Data Sources

To investigate structure vulnerability, we worked with Con Edison engineers to identify sources of raw data from disparate sources within Con Edison, so that we could assemble a single relational database for comprehensive data mining. Our efforts to define and assemble our Consolidated databases have been documented in earlier work [20, 18]. Three major classes of information were found to be relevant for modeling structure vulnerability: structure information (structure type, location and identifier), cable information (including function–meaning main, service or street light, phase or neutral; amount of cable; insulation material; size; year of installation), and history of events. Other features investigated include structure cover type (vented or solid), and inspections data. At the time our project began, some of the data sources were too new to prove useful.[2]

We clean and preprocess tables of structured data. In addition, we utilize a significant source of very noisy textual data known as Emergency Control System (ECS) tickets. These are trouble tickets recorded by Con Edison dispatchers in response to problems in the grid, such as calls from customers about interruptions in service. ECS was instituted in the 1970s, and was expanded in 1986, after Hurricane Gloria. It contains well over 1 million tickets from all boroughs for the period from 1996 through the present. The structured fields contain information such as the ticket date, location, job number, completion time and trouble type; we discuss the trouble type further below. There is also a free text field referred to as the *remarks.* The remarks consist of interleaved lines of free text and structured data, minimally consisting of one or two lines documenting a call from a customer, structured lines indicating who was dispatched to investigate the structure, and actions carried out, such as installation of a temporary shunt (a provisional cable run along the ground) to restore service. Remarks contain many misspellings, acronyms, abbreviations and technical terminology (e.g., CFR for *[cables] cut for replacement*).

Figure 2 illustrates an ECS remarks field. The first line of the ticket explicitly names the category for this ticket, which is SMH (smoking manhole).

An ECS ticket can reference one or more structures, at least one of which will be identified as the source of the trouble. *Information Extraction (IE)* is the process of filling in structured tables or templates with information extracted from textual documents [5]. Through pattern matching by means of the IE functionality of GATE [6], a text engineering tool, we extract fifteen additional attributes. These include structure type (manhole or service box), structure number, metadata indicating whether the event was serious (see next section), terms indicating work was performed on the structure, whether cables were cut for replacement, and similar types of information.

3.2 ECS Events

Each ECS trouble ticket is assigned a trouble type (identified by a three-character mnemonic), out of 288 of possibilities. We worked with our collaborators at Con Edison to select a subset of trouble types relevant to our modeling task. For the problems we investigate in Manhattan, we relied initially on 22 trouble types, and recently expanded this to include 33 trouble types. Two pertain to very serious events: MHX for manhole explosions, and MHF for manhole fires. One pertains to a moderately serious event: SMH for smoking manhole. Of the remainder, 9 are referred

[1] New York City consists of five *boroughs*, or administrative districts: Brooklyn, The Bronx, Manhattan, Queens and Staten Island.

[2] In our current project, we are re-examining existing information and investigating new sources of information, such as construction and repair records, contact voltage detection reports, structure dimensions, weather patterns over time, and yearly volume of salt distributed by New York City Department of Transportation for melting ice and snow.

```
01/21/YR 18:45 FDNY-190 REPORTS A SMH  STREET_1 & STREET_2
01/21/YR 19:35 PERSON REPORTS THE TROUBLE HOLE IS SB-00001
N/W/C STREET_1 & STREET_2......FOUND ON ....SMOKING LIGHTY
01/21/YR 21:55 PERSON REPORTS IN SB-00001 HE FOUND 1 LEG
ON THE 5 WIRE NORTH BURNING IN THE STRUCTURE......CUT/CLEAR
ED & RETIED SAME .......................COMPLETE.............SS
ELIN REPT ADDED FOR INCIDENT:SMH 01/21/YR 22:02 BY PERSON_ID
REPORTED BY: FIRE DEPT
STRUC MSPLATE TYPE NUMBER COND COVTYP COVFOUND DISTANCE
(1) MSPLATE ID SB 00001 WA S Y 00
TYPE OF CURRENT: ALTERNATING CURRENT
VOLTAGE: 120/208V
APPROPRIATE SIZE: 500 MCM
CONDUCTOR CODE: COPPER
POSSIBLE CAUSE OF THE INCIDENT: INSULATION BREAKDOWN
WEATHER CONDITIONS DURING THE INCIDENT: CLEAR
```

Figure 2: Sample ECS Ticket (anonymized): Serious Smoking Manhole

Name	Description	Coefficient	
		Name	Value
Mention	The number of past events where the structure was mentioned in distinct ECS remarks, all years	α_1	0.00170256
RecentMention	Same as Mention, within the past 3 years	α_2	0.00491505
TroubleHole	The number of past events where the structure was the source of the event, all years	α_3	0.00170256
RecentTrHole	Same as TroubleHole, within the past 3 years	α_4	0.00733388
MainPhase	The number of main phase cables in the structure	α_5	0.00114553
ServPhase	The number of service phase cables in the structure	α_6	-0.00019783
Serv_1960_1969	The number of service phase cables installed between 1960 and 1969	α_7	0.00388190

Table 1: Features and coefficients in the ranking model

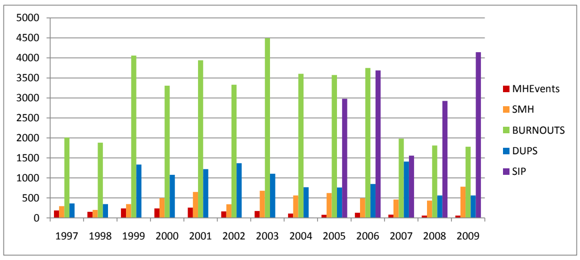

Figure 3: Distribution of ECS tickets across years

to as burnouts, and consist of minor interruptions of service such as SO (side off), WBR (wire burn out), FLT (flickering lights), NL (no lights), NLA (no lights in the area) and so on. Burnouts play a role in our structure ranking model, whereas other trouble types do not.

Identifying the subset of ECS tickets that refer to a distinct event is a critical task in both the structure ranking and the inspections analysis. The mapping between ECS tickets and events is many-to-one. When there is an interruption of service that affects a large building or neighborhood, multiple customers might call in about the same event. Often, a single ECS ticket is identified as the lead ticket, and the other tickets associated with the lead ticket are called *referred* tickets, or *dups*. When we created the 2007 structure ranking, we filtered out all the referred tickets in order to identify a subset corresponding to distinct ECS events.

Figure 3 depicts the distribution of three classes of ECS tickets identified as the lead ticket of an event, manhole events (MHX, MHF), smoking manholes (SMH), the nine burnout ticket types (burnouts), followed by the class of duplicate tickets (any trouble type) and inspection tickets. As shown, manhole events (first bar, in red) are relatively rare, and have been decreasing, smoking manholes (second bar, in orange) have remained relatively constant, and since 1999, burnouts have been decreasing (third bar, in green). Duplicate tickets show a fair amount of variance, but have remained constant on average. The SIP type indicates an inspection ticket; one must be completed for each inspection report. SIPs do not appear until 2005, the first full year of the inspection program.

As described in [18], in an experiment where we concealed the trouble type of ECS tickets, we found that engineers agree only moderately well on a task of assigning a trouble type to ECS remarks. Based on a second pass over the problematic tickets for this task, we developed a high-precision rule-based procedure to classify tickets into three categories for the purpose of labeling structures for supervised ranking. The three categories consist of tickets for serious events, tickets for burnout events, and tickets that do not correspond to distinct events (e.g., referred tickets).

For the inspections analysis reported here, we refined our definition of referred tickets, which reduced the set of lead tickets (the distinct event in a set of referred tickets) by 6.4%. In addition, we identified a new class of tickets whose trouble type should have been SIP, which is the trouble type used for an ECS ticket that is associated with a concurrent inspection report. In some cases, the personnel that should issue an SIP ticket are not able to, and a different unit of Con Edison produces a substitute ticket, using the ACB trouble type. We use two criteria to reclassify certain ACBs as SIPs. When the first lines of the ticket mention a structure type and number, it cannot be a customer call, as customers have no access to these identifiers. If in addition, the same structure has a concurrent inspection report, then we reclassify the ACB ticket as an SIP. Following this procedure to distinguish genuine ACB event tickets from ACB inspection follow-up tickets (pseudo-SIPs), 10.12% of ACB tickets between 2004 and 2009 are reclassified SIPs.

4. STRUCTURE RANKING

Each year, serious events occur on a relatively small proportion of structures; in Manhattan in 2010, there were 60 serious events for over fifty thousand structures. Because of this, we cast our problem within the framework of *rare event prediction*. The goal was to predict whether an event will happen within a given period of time (in our case, a given year), based on what happened before that period. To design each ranking model, we train it to predict events that happened the year before, using data prior to that time. For instance, in order to predict events in 2009, we train the model to predict events in 2008 (knowing what events actually happened that year), using data from 2007 and before.

The ranked list is constructed from a scoring function that gives each structure a real-valued score, where the score indicates its predicted vulnerability. The scoring function is a linear combination of features (covariates) that are derived with respect to the time period for prediction. For instance, one of the features is the number of events that the structure had been involved in within the three years prior to the prediction period. We hand-derived approximated 120 features, all based on our Manhattan secondary grid database. The features came in three major categories: past events, cables, and inspections. The past event features encode the number and types of past events that the structure was involved in over different time periods. The cable features encode the number and types of cables and their installation dates. The inspection features encode the number and results of past inspections, but at this point, have not been powerful enough to use effectively for machine learning; this study is the first indication we have found that the inspections can be useful for prediction, and we are in the process of developing more specialized inspection features based on this study. The label (dependent variable) for a structure is whether the structure was the trouble hole (source structure) for a serious event during the prediction period. We used several different feature selection methods (largest AUC values, information gain, backwards elimination) to choose a small set of features for the model; since serious events are so rare, it is very easy to overfit by including too many features, so the feature selection enables prediction.

We use machine learning techniques for *supervised bipartite ranking* to choose the coefficients for the scoring function, and the structures are then rank-ordered by their score. This algorithm, in combination with the data, determines how important each feature is for prediction. The term *supervised* means simply that there is labeled data to learn from, where the label indicates what the learning model is intended to predict: here, a positive label indicates the structure had a serious event in a given year (was vulnerable), and a negative label indicates the opposite. A brief overview of supervised ranking methods in this context is provided by [22]. The coefficients are chosen (using the training data) to minimize a chosen objective. The objective encodes all positive-negative pairs of examples, and lower values of the objective are achieved when positive examples are ranked above negative examples. The objective is essentially a weighted version of the area under the ROC curve (AUC) [3] that favors the top of the ranked list. The resulting algorithm, which is described in earlier work [19], performs better than others we have tested on the features and labels discussed below, including support vector machine classifiers and pruned decision trees.

Number of inspections	Number of structures	Percentage of structures
1	28,842	56.3111
2	10,226	19.9652
3	4,860	9.4887
4	2,661	5.1953
5	1,639	3.2000
...	...	...
30	8	0.0156
...	...	...
186	1	0.0020

Table 2: Singleton versus multiple inspections

For our first blind prediction test, we aimed to predict serious events in Manhattan during 2007. The seven features shown in Table 1 were selected, and the scoring function is given in Figure 4.

5. INSPECTIONS DATA

The raw data we received from Consolidated Edison includes 126,478 inspection reports in digital format from late 2004 through 2009. From a single raw table of inspection reports containing free text and structured data, we created a normalized relational database model consisting of five inspections tables. As noted earlier, we focus here on the 78,073 inspections that occurred in 2008 and 2009. Of these, 71,890 are on structures with secondary cable (Table 3).[3] During the first years of the program, the inspection procedures were still being refined, and had become relatively stabilized by 2008. The total number of structures in our 2007 Consolidated structures table is 51,219. For just over half the structures (55.31% of structures, 23.19% of inspections), the structure has a single inspection within the 2004-2009 dataset we received. Often, a structure is inspected multiple times, with the maximum being over 100 times. Most of the repeat inspections on a structure are due to the requirement that every time a crew enters a structure, a new inspection report must be completed; maintenance work to reconfigure (*re-rack*) the cables in a structure can require weeks, or sometimes months, to complete. Table 2 shows for values from one to five, for thirty, and for the maximum of 186, the percentage of structures that had exactly that number of inspections. As shown, just over half the structures had a single inspection, another 37.85% had between 2 and 5 inspections.

The first row of Table 3 indicates that for 2008-2009, 34.8% of all inspections on structures with secondary cable

[3] Our dataset includes manholes that have only primary cable.

$$score(structure) = \alpha_1 \times Mentions + \alpha_2 \times RecentMentions + \alpha_{3)} \times TroubleHole + \alpha_4 \times RecentTrHole + \alpha_5 \times MainPhase + \alpha_6 \times ServPhase + \alpha_7 \times Serv\_\mathrm{1960\_1969}$$

Figure 4: Formula for the ranking model

Result of inspections	Count (2008-2009)	Percentage
Clean inspection	25,041	34.83
Level 1 only	17,928	24.94
Level 2 only	1,101	1.53
Level 3 only	18	0.03
Level 4 only	13,234	18.41
Level 1+2	1,417	1.97
Level 1+3	6	0.01
Level 1+4	9,127	12.70
Level 2+3	1	0.00
Level 2+4	1,652	2.30
Level 3+4	33	0.05
Level 1+2+3	2	0.00
Level 1+2+4	2,296	3.19
Level 1+3+4	17	0.02
Level 2+3+4	10	0.01
Level 1+2+3+4	7	0.01
Total	71,890	100.00

Table 3: Inspections outcomes in the 2008-2009 data on structures with secondary cable

resulted in no defects found. The specifications for carrying out inspections sort the remaining types of inspections outcomes into four categories, based on their priority; these are referred to as Levels 1 through 4. The table partitions all inspections into all possible combinations of outcomes. As shown, the largest category after clean inspections (no defects found) consists of structures which had only Level 1 findings. Here we address Level 1 inspection outcomes, where by definition the repair must be completed before leaving the location–except in the event of an emergency, such as a storm, when Level 1 repairs must be carried out within seven days of discovery. Level 2 and 3 inspection outcomes identify issues that can be remediated over two to three years; Level 4 outcomes are for record-keeping purposes only. Because we carried out our analysis before we had access to 2010 data on events, our causal analysis is restricted to the one pair of years for which we had inspections in one year (2008) followed by a year in which we had event data (2009).

6. METHODS

6.1 Causal Inference

The question of whether Level 1 repairs result in a reduced incidence of events involves causal inference. Here, causal inference takes the form of a comparison between structures in a control group that have not received Level 1 repairs, and structures in a treatment group that have. Causal inference requires careful consideration of potentially confounding variables [11]. The outcome of interest here is whether structures that receive treatment (Level 1 repairs) have a lower incidence of events. If the control and treatment structures have distinct outcomes, to infer that it is the treatment that causes the difference in outcomes requires the treatment and control individuals to have an equal likelihood of receiving treatment, all other things being equal (such as confounding variables). If all structures were equally likely to have a Level 1 repair, then it should be possible to identify matched pairs of structures such that for every treated structure, there is at least one similar structure that did not receive the treatment. If treatment structures cannot be matched to control structures, it is not possible to generalize from the treatment to the control.

In an observational study, the process of dividing the data into distinct groups (*strata*) is called post-stratification [23]. Here we use the ranked list to bin structures into groups that have similar rank, where each group is a single stratum. The motivation for relying on structure rank for post-stratification is that a structure will be inspected in one of two ways, one of which we assume to be random, while the other depends on factors closely related to the structure ranking task. Structures are targeted for inspection if they have not had an ad hoc inspection. The latter occur whenever a crew enters a structure, which is often due to an event involving the structure. This means that the likelihood of an ad hoc inspection is related to the likelihood of an event, which in turn is related to our ranked list.

We show in the next section that the treatment and control structures we identify have a high degree of overlap, meaning a similar range of values, and similar distribution of structures across the range of values. We examine overlap of the treatment and control groups with respect to rank, and with respect to the individual features that contribute to the ranking model. The high overlap justifies the inference that the difference in outcomes we find is due to the Level 1 treatment. As we show below, the distributions of treatment and control are similar, but for every treatment structure, there are many matched control structures (i.e., the matching is one-to-many).

6.2 Control versus Treatment

The first PSC-mandated inspections program extended from late 2004 through 2009. During that period, the inspection procedures evolved, were standardized in 2008, and fully implemented by 2009. At the time we carried out our analysis (the fall of 2010), the last year for which we had a complete record of events was 2009. Given the evolving inspections procedures, and the short time frame for looking into the future, our observational analysis of a treatment effect focuses on 2009 events relative to 2008 inspections.

We binned the treatment and control structures into eight categories, based on feedback from our collaborators at Con Edison about the desired granularity of the analysis. The most vulnerable category consisted of the top 5,000 structures in the ranked list. Due to a mix of opinions from Con Edison as to whether we should rely directly on the ranked list, we defined the next seven categories in terms of features that contributed to the ranking model, rather than by the ranking itself. The average number of structures in each category was approximately 7,000. For this paper, we defined all eight categories directly in terms of the ranking.

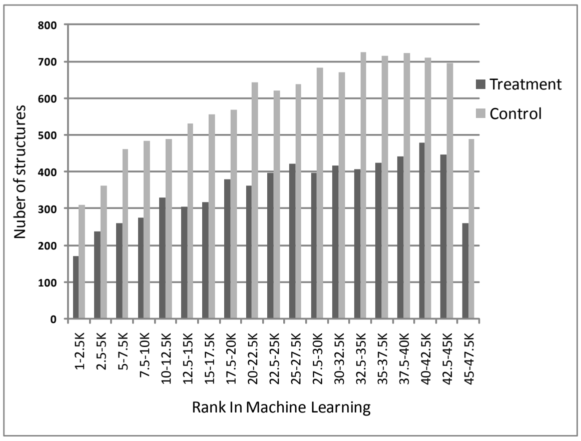

Figure 5: Treatment and control groups have good overlap: structures in both groups for the full range of values, and with similar proportions

Category 1 is again the top 5,000 structures, and each next category is the next 7,000 structures in the ranked list, with somewhat less than 5,000 structures in the final eighth category.

For each bin, the control group is a group of structures that has had no inspections through 2008. The treatment group has had no inspections before 2008, and in 2008, all inspections on the structures in this group have Level 1 findings, no findings for Levels other than Level 1, and no clean inspections. In thsi manner, the control group is ensured to have no inspection-triggered repairs, and no clean inspection reports through 2008. The treatment group is ensured to have no clean inspection reports through 2008, and no inspection-triggered repairs until 2008. In 2008, the treatment group had Level 1 repairs only, and no other types of inspection findings. In sum, the only difference between the two groups is that the treatment group had a Level 1 repair in 2008.

After defining the treatment and control groups, we then look at the incidence of events on these structures in 2009. In this way, we can test whether the treatment group (Level 1 repairs) has a lower incidence of 2009 events.

To determine whether the use of structure rank to match control and treatment structures is well-motivated, we assess their overlap, meaning whether the control and treatment structures have similar balance and distribution. The two groups are balanced if they have the same range of values, and they have equivalent distributions if the same proportion of control group and treatment group occur at each value (or interval of values). Figure 5 shows the overlap (similarity in balance and distribution) for total number of cables, which is one of the features that contributes to the ranking model (see Section 4). The same overlap holds true for the remaining features in the ranking model.

7. RESULTS

To test for a treatment effect within each of the eight vulnerability categories, we generate 2 × 2 contingency tables for each category of *Treatment vs. Control* by *Structures with 2009 Events (+Events) vs. Structures with no 2009 Events (-Events)*. We use Fisher's Exact Test [8] to test for significance. Table 4 shows the contingency table for Category 1 structures. The results for the full data, and for each category, are summarized in Table 5, which gives the cell values for each contingency table, along with the p-values for Fisher's Exact Test.

	+ Events	− Events	Row Marginals
Treatment	52	377	429
Control	114	592	706
Column Marginals	166	969	1135

p=0.0370

Table 4: Contingency table for Category 1 structures

Figure 6 summarizes the findings. The x-axis shows the eight categories, in order of decreasing vulnerability. The y-axis indicates the proportion of treatment (solid line) or control (dashed line) structures in that category that were associated with events, either as the trouble hole of the event, or mentioned in the ticket along with the trouble hole. Figure 6 shows that for all categories apart from the one with the lowest vulnerability, the treatment group that received Level 1 repairs has a lower incidence of 2009 events compared with the control group. The p-values shown in Table 5 indicate that the differences are highly significant for the case comparing all treatment structures to all control structures (N=19,151), and for Categories 1, and 4 through 7.[4] Second, each next vulnerability category has a lower incidence of events in both the treatment and control groups. Both findings demonstrate the benefit of applying the Secondary structure ranking model for Manhattan to the analysis of the inspections data.

8. CONCLUSION

Prior to the task of applying the ranked list to the analysis of inspections, the ranked list of structures we could produce for each borough had been considered a component in prioritizing repairs and upgrades, such as replacement of service

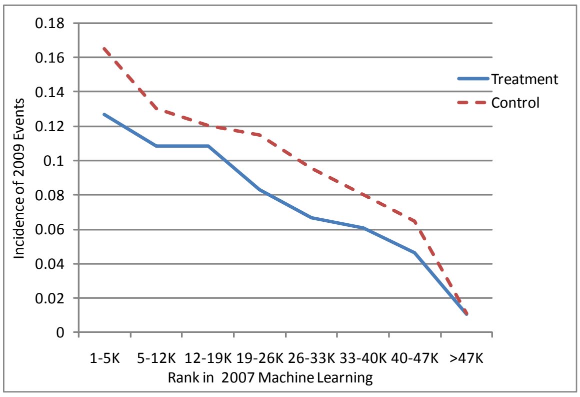

Figure 6: Incidence of 2009 events for 8 categories of treatment vs. control structures

[4] An analysis we delivered to Con Edison used the ranked list only for Category 1, and features from the ranking model for all subsequent categories. In that analysis, the differences between treatment and control were statistically significant for Categories 1-6, and not for Categories 7 and 8.

	T	C	T	C	
Cat	+ Evt	+ Evt	− Evt	− Evt	p-value
All	532	1152	6486	10981	3.055×10^{-6}
1	52	114	377	592	0.0370
2	88	181	690	1184	0.1067
3	105	191	841	1350	0.1830
4	92	210	969	1537	0.0030
5	72	172	1071	1752	0.0050
6	67	158	1161	1749	0.0036
7	53	114	1168	1847	0.0407
8	3	12	254	970	0.6203

Table 5: Contingency table cell values for the full dataset, and for each category, with p-values

box covers. Here it was applied in a retrospective analysis of data collected during application of a program to inspect all structures. In a real sense, all data mining applications address sustainability by leveraging a resource, namely existing data, in novel ways. Our work applied causal inference to an existing set of inspections reports in a fashion that demonstrated a new utility for our structure ranking work. By showing that more vulnerable structures are indeed more susceptible to future events, we also inspired Consolidated Edison to consider a new approach to optimizing the order in which secondary structures are inspected, a problem our future work will address.

Because the eight categories of structures in Figure 6 are defined solely with respect to the ranking model, the smooth downward slope of the plots provides confirmation that the ranked list represents vulnerability of structures to events. Each next category has a lower proportion of its structures that experiences an event of any type in 2009. This in turn demonstrates that it is possible to apply data mining and machine learning methods to real world data that were never intended to support predictive or causal inference.

For Con Edison, the benefit of the work described here is the ability to quantify the impact of its programs. The inspections program is only one of several programs designed to improve the safety and reliability of the distribution network. While it had been observed that the incidence of secondary events had decreased, the work presented here isolates the contribution of a specific component of the inspections (the high priority Level 1 repairs), quantifies the reduction in events relative to the vulnerability of sets of structures, and points to which sets of structures have a reduced incidence that, given the statistical evidence, is a result of Level 1 repairs.

9. ACKNOWLEDGMENTS

We thank our co-authors at Con Edison for the opportunity to work with them on the secondary data, and for identifying the inspections data as a potential application of our ranked list. We thank them and colleagues on earlier work at Con Edison, including Delfina Isaac and Graciela Varela-Maloney, for generous amounts of time and expertise.

10. ADDITIONAL AUTHORS

Additional authors: Steve Hanebuth and Steve Ierome and Debbie Pangsrivinij (Consolidated Edison Company of New York, email:
(HanebuthS|IeromeS|Pangsrivinij)@coned.com)

11. REFERENCES

[1] S. Agarwal, D. Dugar, and S. Sengupta. Ranking chemical structures for drug discovery: A new machine learning approach. *Journal of Chemical Information and Modeling*, 50(5):716–731, 2010.

[2] H. Becker and M. Arias. Real-time ranking with concept drift using expert advice. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD '07)*, pages 86–94, New York, NY, USA, 2007. ACM.

[3] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[4] G. Chen and A. T. Peterson. Prioritization of areas in China for the conservation of endangered birds using modelled geographical distributions. *Bird Conservation International*, 12:197–209, 2002.

[5] J. Cowie and Y. Wilks. Information extraction. In R. Dale, H. Moisl, and H. Somers, editors, *Handbook of Natural Language Processing*, pages 241–69. Marcel Dekker, New York, 2000.

[6] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL '02)*, pages 168–75, 2002.

[7] M. Dudík, S. J. Phillips, and R. E. Schapire. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research*, 8:1217–1260, Jun 2007.

[8] R. A. Fisher. *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh, UK, 1925.

[9] Y. Freund and R. E.Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55, 1997.

[10] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:993–69, 2003.

[11] A. Gelman and J. Hill. *Data Analysis using Regression and MultilevelHierarchical Models*. Cambridge University Press, Cambridge, UK, 2006.

[12] P. Gross, A. Boulanger, M. Arias, D. L. Waltz, P. M. Long, C. Lawson, R. Anderson, M. Koenig, M. Mastrocinque, W. Fairechio, J. A. Johnson, S. Lee, F. Doherty, and A. Kressner. Predicting electricity distribution feeder failures using machine learning susceptibility analysis. In *The Eighteenth Conference on Innovative Applications of Artificial Intelligence IAAI-06*, Boston, Massachusetts, 2006.

[13] R. Jiang, H. Yang, L. Zhou, C.-C. J. Kuo, F. Sun, and T. Chen. Sequence-based prioritization of nonsynonymous single-nucleotide polymorphisms for the study of disease mutations. *Am J Hum Genet.*, 81(2):346–360, aug 2007.

[14] J. Kirtley Jr., W. Hagman, B. Lesieutre, M. Boyd, E. Warren, H. Chou, and R. Tabors. Monitoring the health of power transformers. *IEEE Computer Applications in Power*, 9(1):18–2, Jan 1996.

[15] R. J. Landman. Underground secondary AC networks, a brief history, August 4 2007. Presented at 2007 IEEE Conference on the History of Electric Power.

[16] P. Mirowski, Y. LeCun, D. Madhavan, and R. Kuzniecky. Comparing svm and convolutional networks for epileptic seizure prediction from intracranial eeg. In *Machine Learning for Signal Processing, 2008. MLSP 2008. IEEE Workshop on*, pages 244 –249, oct. 2008.

[17] J. F. Murray, G. F. Hughes, and K. Kreutz-Delgado. Machine learning methods for predicting failures in hard drives: A multiple-instance application. *Journal of Machine Learning Research*, 6:783–81, 2005.

[18] R. Passonneau, C. Rudin, A. Radeva, and Z. A. Liu. Reducing noise in labels and features for a real world dataset: Application of NLP corpus annotation method. In *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, 2009.

[19] C. Rudin. The P-Norm Push: A simple convex ranking algorithm that concentrates at the top of the list. *Journal of Machine Learning Research*, 10:2233–2271, 2009.

[20] C. Rudin, R. Passonneau, A. Radeva, H. Dutta, S. Ierome, and D. Isaac. A process for predicting manhole events in Manhattan. *Machine Learning*, 80:1–31, 2010.

[21] C. Rudin and R. E. Schapire. Margin-based ranking and an equivalence between AdaBoost and RankBoost. *Journal of Machine Learning Research*, 10:2193–2232, 2008.

[22] C. Rudin, D. Waltz, R. N. Anderson, A. Boulanger, A. Salleb-Aouissi, M. Chow, H. Dutta, P. Gross, B. Huang, S. Ierome, D. Isaac, A. Kressner, R. J. Passonneau, A. Radeva, and L. Wu. Machine learning for the New York City power grid. *IEEE Transactions on Pattern Analysis and Machine Intelligence, To Appear*, 2011.

[23] T. M. F. Smith. Post-stratification. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 40(3):315–23, 1991. Special Issue: Survey Design, Methodology and Analysis (2).

[24] J. Steed. Condition monitoring applied to power transformers-an rec view. In *Second International Conference on the Reliability of Transmission and Distribution Equipment*, pages 109–11, 1995.