21st-Century Data Miners Meet 19th-Century Electrical Cables

Cynthia Rudin, Rebecca J. Passonneau, and Axinia Radeva, *Columbia University*
Steve Ierome and Delfina F. Isaac, *Consolidated Edison of New York*

Researchers can repurpose even extremely raw historical data for use in prediction.

Electrical grid reliability will be a key issue as peak demand for electricity continues to increase. Grids will need to accommodate a growing population, more high-tech appliances, distributed power generation, and, soon, a large fleet of electric vehicles.

Considering the future of smart grids in the world's major urban centers, it's tempting to picture completely remade grids of shiny new copper wires, each implanted with smart status monitors. This is very unlikely, however, because the existing electrical infrastructure is enormous. New York City alone has more than 94,000 miles of underground cable, enough to wrap around the Earth three and a half times. There is simply too much cable to replace or individually monitor—we're not even close to having truly smart grids.

Are there ways we can make an energy grid smarter without using monitors to assist with reliability? Perhaps there are ways we can learn from reliability failures in the past in order to help maintain reliability and safety in the future. In a sense, can we build a "historically conscious" smarter grid? Our team of scientists at Columbia University and engineers at Consolidated Edison (Con Edison), New York City's power utility, set out to answer a version of this question.

Specifically, we sought to determine whether Con Edison data regarding past failures on the city's low-voltage grid—manhole fires, explosions, smoking manholes, and burnouts—could be used to predict, and thus prevent, future events. Our goal was to create a list of manholes (which serve as access points to the underground electrical grid), ranked from most to least vulnerable, that could support the company's inspection and repair programs, which improve public safety and energy reliability.

PROCESSING RAW DATA

Many electrical grids are very old, and cables from the Thomas Edison era are still reliably carrying current. New York City has the world's oldest grid, and we computed that over 5 percent of Manhattan's low-voltage underground cables were installed before 1930. Con Edison's databanks started in the 1880s, but certainly those historical data weren't created for the purpose of predictive modeling. The historical data are extremely raw and very noisy.

The tasks of matching up failure events to the manholes they refer to and matching manholes to the cables they contain are complicated, mainly because of the various means by which Con Edison has recorded data over the years. Our cable data comes from the company's accounting department, past event records come from the emergency control systems database, and manhole location, inspection, and other data come from different Con Edison sources.

We're trying to predict "serious events," but when viewed through the lens of historical records, it's not always clear whether a past event was serious. Con Edison records events through "trouble tickets." These shorthand notes taken by dispatchers

0018-9162/11/$26.00 © 2011 IEEE Published by the IEEE Computer Society

```
FIRE DEPT.REPORTS;CONDITION ORANGE F/O 1655 WEST END AVE.
01/26/00 11:54 MDEPICA DISPATCHED BY 71122
01/26/00 12:19 FERRARO REPORTS: REL. FD. TBL/H M-493784
F/O 1655 WEST END AVE... FOUND RD SOLID COVER 3' OFF AGAINST
VEHICLE...HOLE IS SMOKING...REQS..FLUSH.................
CO = 0PPM -> 1655 WEST END AVE...BASEM'T AREA............DR
01/26/00 14:36 FERRARO REPORTS: IN M-493784 F/O 1655 WEST ST.
HAVE (1)3-500,2-4/0,COPPERED GOING TO BUS COMP. V-72184...DR
01/26/00 18:00 PICA REPORTS: ========= C.F.R.===============
FROM M-493784 F/O 1655 WEST END AVE.. 3-500,2-4/0,AC,4'53'
TO BUS COMPARTMENT V-72184 F/O 1655 WEST END AVE..........
ALL B/O CLEARED.......... ALL CO CLEARED...............DR
01/26/00 18:00 MDEPICA COMPLETE BY 23349
***************ELIN REPORT MADE OUT**********MC
07/29/00 00:20 ACT TRBL CHNGD FROM EDSMHF TO EDSMHX BY 71453
08/01/01 09:54 REFERRED TO: CAI ES0012 FYI BY 01585
```

Figure 1. Excerpt of a Con Edison trouble ticket (edited for anonymity) for a manhole explosion event in 2000.

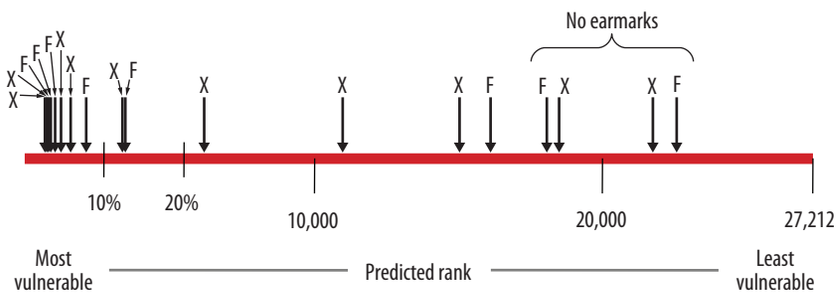

Figure 2. Results from the Bronx blind prediction test. The axis is the predicted rank. Each arrow labeled "X" indicates a 2009 manhole explosion, and each arrow labeled "F" indicates a 2009 manhole fire.

describe the company's responses to events, not the events themselves, and are difficult for a nonexpert to decipher. Figure 1 shows an excerpt of a trouble ticket (edited for anonymity) for a manhole explosion event in 2000.

The text in trouble tickets is very irregular and thus challenging to process in its raw form. There are many spellings of each word—for instance, the term "service box" has at least 38 variations including SB, S, S/B, S.B, S?B, S.B., SBX, S/BX, SB/X, S/XB, /SBX, S.BX, S &BX, S?BX, S BX, S/B/X, S BOX, SVBX, SERV BX, SERV-BOX, SERV/BOX, and SERVICE BOX.

To determine whether an event is serious, we extract information from the ticket text. For example, variations of terms like "smoking" likely connote a serious event, while cable sizes like "2-4/0" or "3-500" or the term "cleared" indicate that an event occurred, since a cable replacement was required.

After processing the tickets, we combine them with the processed cable and manhole location data and several other data sources to obtain an accurate reconstruction of each manhole's decade-long event history, potentially 120-year-old cable history, and inspection results. We then use machine learning algorithms to create a meaningful event prediction model, targeted to predicting failures on individual manholes.

The data processing, and how it's coupled to the statistical model, is the most important step within our knowledge discovery process. Our approach doesn't transform raw data into a form that can be used for many different tasks; generic data transformations wouldn't suffice. Instead, the data processing is specifically geared to our specific modeling task and is driven by the goal of building the predictive model.

Our data processing was done in a transparent way, to ensure that we could explain to Con Edison engineers and managers why a particular manhole was ranked highly, point to the past tickets it was involved in, and describe why we might recommend it for repair or inspection.

PREDICTING SERIOUS EVENTS

We conducted blind tests of the event prediction model in three New York City boroughs: Manhattan, Brooklyn, and the Bronx. For these blind tests, the goal was to predict serious events that happened after the current end date of our database. For instance, when we have data through the end of 2008, we try to predict events in 2009.

Figure 2 shows results from the Bronx test, in which we aimed to predict 2009 events using complete data through 2007 and partial data from 2008. The axis is the predicted rank, from most vulnerable to least vulnerable. Each arrow labeled "X" indicates the predicted rank of a 2009 manhole explosion, and each arrow

A NEW FOCUS

With this issue, *Computer* replaces AI Redux with a new column on the use of analytics, data mining, and machine learning to foster discovery in diverse domains. The goal of the Discovery Analytics column is to highlight the pervasive role these technologies now play in science, engineering, humanities, and beyond. Send comments and future article suggestions or proposals to column editor Naren Ramakrishnan, Department of Computer Science, Virginia Tech, Blacksburg, VA; naren@cs.vt.edu.

labeled "F" indicates a 2009 manhole fire. The top 10 percent of manholes (2,721/27,212) on our ranked list contained 44 percent (8/18) of the manholes that experienced a serious event in 2009; the top 20 percent of manholes (5,442/27,212) on the list contained 55 percent (10/18) of the manholes that had a 2009 serious event.

As part of the blind testing, we did case studies of the manholes that experienced a serious event but were at the bottom of the list to understand why or whether our model had failed in those cases. For the case studies, we relied on tools designed to facilitate communication between the Columbia scientists and Con Edison engineers.

One of the tools summarizes all of the raw and processed data that goes into the model regarding a given manhole. This "report card" tool allows us to tell at a glance precisely which trouble tickets, cable records, and other data determine the manhole's ranking. The reports on the bottom four manholes in Figure 2 (labeled "no earmarks") gave no indication that they were vulnerable—these manholes had very few cables and no involvement in past events. Even with a clean and comprehensive database, manhole event prediction can sometimes be a difficult task.

Another tool displays the underground electrical grid's geometry superimposed on Google Earth satellite images, as Figure 3 shows. Each circle in the image is a manhole, and each line between two manholes represents underground low-voltage cables connecting them. This visualization tool enables us to check that our match of cables to manholes is correct and complete and shows the density of trouble tickets in particular neighborhoods.

The Columbia-Con Edison manhole event prediction project provides important lessons.

Figure 3. Image from a visualization tool that displays the underground electrical grid's geometry superimposed on Google Earth satellite images. Each circle is a manhole, and each line between two manholes represents underground low-voltage cables connecting them.

First, researchers can repurpose extremely raw historical data for use in prediction. Databanks similar to Con Edison's are commonly not repurposed, left instead to become "data tombs." But researchers often can analyze and exploit such data to make important contributions—in this case, to devise a better procedure for electrical grid inspection and repair that could improve public safety and energy reliability. The challenge is navigating an ocean of possible data processing tasks, some more rewarding than others, to achieve a more accurate predictive model.

Second, it's possible to maintain future electrical grids using past data. The backbone of our future's smart grids will, for some time, be what exists now. Many power companies face similar challenges, but the usefulness of their historical data will depend on how they collect and store the data, and whether they're willing to expend the resources (and take the risk) to mine it. Our results demonstrate that the investment can be well worth the risk, in that it can help make the grid smarter, safer, and more reliable.

Cynthia Rudin *is an assistant professor in the Operations Research and Statistics group at the MIT Sloan School of Management as well as an adjunct research scientist at Columbia University's Center for Computational Learning Systems. Contact her at rudin@mit.edu.*

Rebecca J. Passonneau *is a senior research scientist at Columbia University's Center for Computational Learning Systems. Contact her at becky@ccls.columbia.edu.*

Axinia Radeva *is a staff associate at Columbia University's Center for Computational Learning Systems. Contact her at axinia@ccls.columbia.edu.*

Steve Ierome *is a distribution engineering secondary system analysis manager at Consolidated Edison of New York. Contact him at ieromes@coned.com.*

Delfina F. Isaac *is a quality assurance manager at Consolidated Edison of New York. Contact her at isaacd@coned.com.*