Energy and Environmental Statistics
G8325

Lauren Hannah

Columbia University, Department of Statistics

September 4, 2013
First, let’s get to know each other.

- Name
- Affiliation
- Research interests
- Why you are taking this course
Now, a bit more about the course.

▶ What is the idea behind this course?
Now, a bit more about the course.

- What is the idea behind this course?
- What are the basic problems?
Now, a bit more about the course.

- What is the idea behind this course?
- What are the basic problems?
- Why are these problems hard to solve?
What material will this course cover?

- **Decision making**: stochastic optimization and methods for sequential decision processes
- **Modeling**: Bayesian nonparametric methods including Gaussian processes and Dirichlet processes, with inference to scale those methods to massive datasets
- **Selected topics**: topics associated with projects (possible topics include missing/corrupted data, spatio-temporal statistics, causality, co-clustering, dimension reduction, aggregation methods, etc)
What do I expect from you?

- Class participation
  - Classes are (mostly) discussions about readings. Be prepared to discuss!
  - Prepare one reading for discussion with the rest of the class.
  - Scribe for one class (so notes can be posted online).
  - Reading responses on all readings. This is 1 paragraph to 1 page about the readings. What did you not understand, what did you learn, what can you apply to your research? Email them to me the night before class.
  - Short presentations on your problems at the beginning of the semester.
  - Longer presentations on your problems at the end of the semester.
What do I expect from you? (continued)

▶ Project
  ▶ Pick a project by week 3
  ▶ Give a short (2-5 min) presentation on what questions the dataset can answer and what challenges it presents
  ▶ Project proposal by week 5
  ▶ Midterm report by week 10
  ▶ Final report by week 14

▶ Topic suggestions: what do you want/need to learn?
Let’s go over the syllabus in detail and pick presentation weeks....
The Columbia Center for Computational and Learning Systems (CCLS) is working with a power management company, Rudin, to control the systems (electric, steam, etc) for 8 Manhattan office buildings. They have:

- occupation data
- electricity usage
- temperature (interior and exterior)
- steam usage
- ...and a whole bunch of questions
Preheating... or not?

Several of the office buildings use steam heat. Heating bills are composed of:

- standard cost per megabar (Mb) used
- a super peak charge, that is $\sim$1,100 times the standard cost for the highest 15 minute usage in the billing period during peak hours

When should we preheat? (What data do we have? What makes this problem difficult?)
Preheating... or not?

When should we preheat? (What data do we have? What makes this problem difficult?)

Steam Usage Comparison

When should we preheat? (What data do we have? What makes this problem difficult?)
Each of these buildings has thermostats, with different possible set points for each level. We want to minimize electricity costs within a set of constraints:

- should be above minimum temperature
- should be below maximum temperature

The data:

- observed set points
- observed energy usage
- observed temperatures

What should the set points be? (What makes this problem difficult?)
CCLS has (varying amounts of) data for 8 buildings. We would like to group the buildings based on:

- thermodynamics
- HVAC system
- occupancy size
- ...

How should the buildings be clustered? Latent variables and mixed membership? How can these clusters be used to generalize models?
Start-up Recommendation

When should building managers start the building systems for the day?

- currently, based on what was done in “similar” days using SVM
- gives time to turn everything on at once
- the recommendations are probably suboptimal
- have occupancy, weather data
- want to minimize electricity and steam costs subject to temperature constraints

What should the policy be? What are the challenges here? Do we have enough data?
Ramp-down Recommendation

When should building managers turn off the building systems for the day?

- currently, based on what was done in “similar” days using SVM
- gives time to turn everything off at once
- the recommendations are probably suboptimal
- have occupancy, weather data
- want to minimize electricity and steam costs subject to temperature constraints

What should the policy be? What are the challenges here? Do we have enough data?
Some of the buildings are heated by steam, but there is a secondary water system to heat the air in the building perimeter. Currently, managers follow a specific schedule based on outside air temperature. The schedule is based on simulation models made by engineers, rather than data. Can we make a recommendation based on data to reduce energy usage and cost?
A common operational procedure available on a BMS is to close all outside air dampers and to recirculate air that is already within the building. This is done on especially hot and cool days to delay energy-intensive cooling/heating in the morning. The constraints are:

- this raises CO₂ levels, which must be kept below a certain threshold
- this may reduce energy usage, and thus cost

What policy recommendations should be made? What data would be used to support these recommendations?
The speed setting on a chiller determines the rate at which the chiller produce chilled water at a specific setpoint to be supplied to the fan. However:

- if the chiller speed is set too low or too high for the amount and temperature of water, then the chiller may trip, surge or run inefficiently
- data-based speed recommendation may save energy

What should the policy recommendation be? What are the difficulties for this problem?
Currently, there is a two hour ahead space temperature forecast based using SVMs. Main goals are:

- error reduction
- increased stability of estimate

What are better prediction methods? How can we evaluate them?
Anomaly Detection

From time to time, there are unusually high or low values recorded in our data. Notable variables that exhibit such anomalies include energy usage, temperature readings, and HVAC equipment speeds and statuses. This could be caused by:

▶ deliberate changes in managers’ actions
▶ outside events like weather or unusual circumstances
▶ significant system problems
▶ poor model behaviour
▶ good, old-fashioned bad data

We would like to find and categorize anomalies. What are the difficulties? What are the opportunities?
Jessica, this one is on you to explain.
Data to determine when fans and pumps turned on or off is sparse. Building managers turn on fans and pumps in staggered sets to prevent overloading the system. They also run experiments during the night to better understand building behavior. Additionally, building managers turn on the pumps to prevent the pipes from freezing. We have data that is affected by fan and pump status such as steam usage and space temperatures. Using the data available, we'd like to determine the latent behaviors of building managers turning the fans and pumps on and off. This information will improve the quality of our data and help improve our other models.
Freecooling

It’s a nice day out... so open the windows!

- outside temp is 55-65 with no heating/cooling needed
- mix outside air with recirculated air in varying ratios
- saves a lot of energy
- ...but building managers have little control over space temperatures

When should this be employed? What effect would it have on building temperatures? And how much money would it save?
Filtering Bad Data

Bad data points are frequent (and unlabeled!) in BMS data. The current system uses all data, with only limited, manual filtering. What are the difficulties? What are the opportunities?
Many buildings have two or more chillers to accommodate high heat-load days. The increase in cooling potential and energy use with the addition of operating a second chiller is significant, but if there is not significant heat load impacting the building than a second (or third) chiller operating will lead to overcooling. We would like to devise a system to determining how many chillers are necessary to supply comfortable air to building spaces in order to increase the efficiency of the cooling system and decrease instances of over-cooling.
Environmental Datasets.
Data for sea surface temperature:

- monthly since 1864
- weekly since 1981
Sea Surface Temperature

Weekly Average SST
2011/01/30 - 2011/02/05

NOAA/ESRL/PSD
Data issues:

- data are sparse
- data are missing (early years, world wars, etc)
- not collected in ice cap areas
- data collection methods have changed
- some measurement errors
Possible Problems

- get a low dimensional representation of the data (interpretable!), both worldwide and regional (south Atlantic, tropical Pacific, etc)
- better long term forecasts of phenomena like El Nino
- combine existing models/forecasts to get better predictions of El Nino
- denoise data/correct for measurement bias
- is there a causal effect between CO$_2$ concentration and average SST? Strength of El Nino? Etc.
Combining Datasets

There are other datasets that can be attached to the SST data:

- multi-satellite blended ocean winds
- historical temperature anomalies
- historical precipitation anomalies

Can these be used to make better prediction models? Interpretable sparse models?
Ensemble Models

Many phenomena are predicted/summarized by a collection of models to express uncertainty. Model combination is fairly crude and usually does not account for external factors.

- Is there a way to use the ensemble for a better point estimate?
- What about a distribution?
- How can features/covariates be selected or included within a new model?

There is often limited data (maybe monthly data since the late 1980’s. How can these be effectively tested? What about the time series nature of the data?
We have tornado data since the 1950’s.

- observational (all observed tornadoes recorded)
- time stamp
- touchdown longitude/latitude
- end longitude/latitude
- monetary damage
- casualties
- F class
Tornadoes

Lauren Hannah
Energy and Environment
September 4, 2013 35 / 38
Data problems:

- counts change over time
- highly overdispersed
- tornadoes are attracted to major highways
- data might be fudged to fit forecast
Tornadoes

Possible problems:

▶ hazard rate
▶ does “Tornado Alley” exist?
▶ ways to incorporate topography?
▶ is there a change in the rate of tornadoes?
▶ are the expected losses growing?
Next Classes

Next week:
▶ readings from Spall, posted on class website
▶ reading response due day before class

In two weeks:
▶ pick a problem/dataset
▶ be prepared to discuss dataset including challenges, and which problem you would like to solve with that dataset