Enhancing Sustainability: Building Modeling Through Text Analytics

Tony Kassekert, The George Washington University

Jessica N. Terman, George Mason University

Research Background

- Recent work by Terman et. al (2015) founds the role of grant management significantly impacts sustainable policy implementation delays.
- Extant research on federal federalism indicates that goal congruence improves performance (Nicholson-Crotty 2004).
- Several theoretical inconsistencies with previous literature occurred when our team tried to combine both set of hypotheses in a single model.
  - The purpose of this presentation is to explain how we are using text mining to improve our estimation and achieve theoretical consistency.

Research Question and Hypothesis

- Does economic development as a motivation for sustainable development impact implementation?
  - Are local managers more technocratic? Do they more consistently work on implementation?
  - Economic development motivations in particular would indicate a preference for spending funds on the grant in a timely fashion.

- **We hypothesize** that when local governments are focused on sustainability as an economic development tool, they are more likely to complete projects on time for similar projects.

Data Sources and Methods

- Data:
    - Department of Energy administrative data
        - All grant application text is directly from DOE.
    - National survey to the population of EECBG grantees
        - Over 50% response rate in 2009
    - Census bureau

- Methods:
    - Bayesian clustering of textual data (tm and bclust R package)
    - A relative risk survival (log-binomial) model is used to estimate implementation delay times.
        - The model has a robust error based on the cluster.

Variables and Measurement

- Dependent variable (days of delay):
  - Delay between jurisdiction-proposed EECBG start date (when they receive funds) and the actual date of funds dispersal/use.
  - A positive coefficient equates to longer delays, while a negative coefficient indicates less delay.
- Independent variable
  - The variable of interest is economic development motivation
    - This is a factor score created off from several survey questions.

All Independent Variables

Satisfaction w/DOE Application Process	Citizen advocacy level
Satisfaction w/DOE Approval Process	Number of prior sustainable policies
Satisfaction w/DOE Tech. Support	Green practices count
Administrative Capacity	Green development in planning
External application assistance	Economic development tool
Citizen application participation	Budget (logged)
Copied policies from other governments	Unemployment
Innovative (new) policies to implement	Manager Form of Government

- These were all chosen to compare results to a previous paper from APPAM.

Research Problem

- Most survey research classifies projects into groups that are not mutually exclusive.

- For example, the EECBG grant process had localities pick between:

  Energy Efficiency Strategy
  Technical Consultant Services
  Buildings Audits
  Financial Incentive Program
  Energy Efficiency Retrofits
  Buildings and Facilities
  Transportation
  Codes and Inspections
  Energy Distribution
  Material Conservation Program
  Reduction Greenhouse Gases
  Lighting
  Onsite Renewable Technology
  Other

- For example, LED lighting could be in the lighting category or part of a building retrofit, or energy efficiency strategy.
  - This measurement error creates inefficacies in estimate standard errors.

Text Mining as a Solution

- We propose text mining energy grant proposals to augment survey data and administrative records.

- Using text analytics, we can classify the grants by their text in an effort to determine which proposals were more similar.

- This allows us to cluster similar projects in a more accurate manner without the unnecessary measurement error.

Short Review of Text Mining

- Text mining is analogous to other exploratory statistical techniques.
  - The primary method for both is cluster analysis.
  - A second frequently used tool for text is singular value decomposition (SVD) is similar to principal components analysis.

- Text mining basically develops a numeric representation of the textual data and analyzes it with standard tools.
  - The standard approach treats documents as rows and terms (e.g. words) as columns.
  - This creates a very large, sparse matrix (lots of zeros)

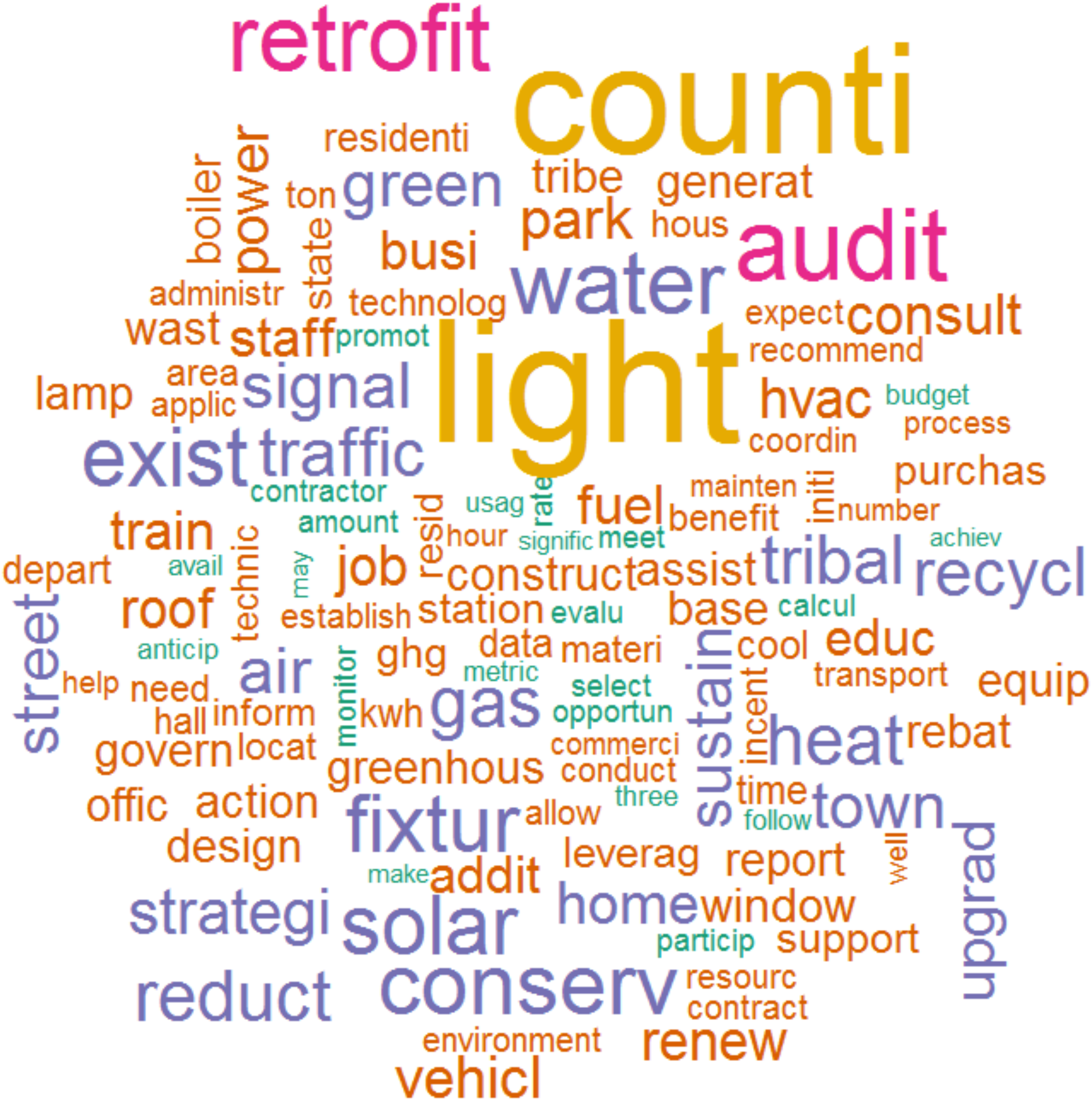
- Text mining is not the same as data mining, although the two are often used in concert.

Transforming Text to Usable Data

- There are several steps that are commonly applied:
  - Normalize case (make everything lowercase)
  - Remove punctuation
  - Remove white space
  - Remove numbers
  - Remove stopwords (the, in, a…)
  - Stem words (chop off end of words- ing, es, er)
    - This means finding the core of a word (city = cities)
  - Choose a weighting scheme
    - Often we weight words to adjust for frequencies and/or document length.
    - The Euclidian distance (used in both clustering and factor analysis) is often not the best choice for text.
- Reduce dimensionality

Reducing Dimensionality

- Text mining suffers from too much information, so we want to reduce it down to something manageable.
- Words that appear in all the texts are useless at discriminating between them.
  - For example, the word energy appears in every grant proposal, so it adds no value to choosing between them.
    - Think of a regression variable that equals 1 for 98% of your cases and 0 otherwise. It would be unlikely to be predictive.
  - Text mining usually begins by removing these too frequently occurring words.
- Words that appear in too few texts add little value also.
  - Not frequent enough to compare between groups.
    - Think of a regression variable that equals 1 for 3% of your cases and 0 otherwise. It would be unlikely to be predictive.

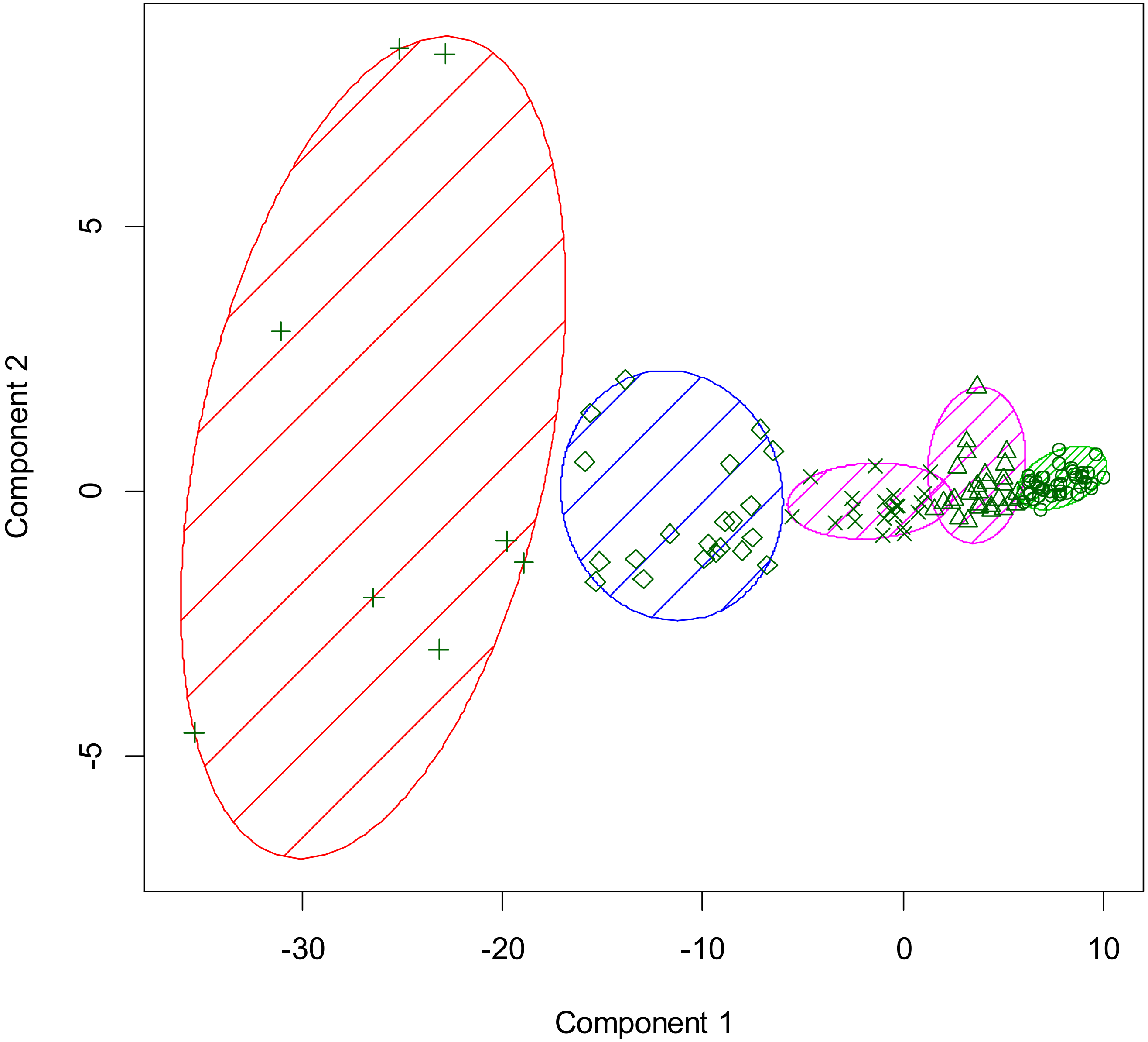

retrofit
counti
residenti
boiler
power
ton
green
tribe
generat
state
park
hous
busi
audit
water
administr
technolog
expect
consult
wast
staff
promot
light
recommend
area
lamp
applic
signal
hvac
budget
process
coordin
exist
traffic
purchas
contractor
usag
rate
fuel
mainten
initi
amount
resid
hour
signific
meet
benefit
number
achiev
train
depart
avail
technic
may
job
construct
assist
tribal
recycl
street
roof
establish
station
evalu
base
calcul
anticip
air
ghg
data
materi
cool
educ
help
need
monitor
metric
select
sustain
transport
equip
hall
inform
kwh
gas
opportun
incent
heat
rebat
govern
locat
greenhous
commerci
conduct
offic
action
fixtur
allow
three
time
follow
town
design
well
upgrad
make
addit
leverag
report
strategi
solar
home
window
particip
support
reduct
conserv
resourc
contract
environment
renew
vehicl

Text analysis

- After all the data cleansing, the document term matrix (dtm) is analyzed.

- For purposes of this presentation, we will cluster the texts.
  - We use a Bayesian clustering
  - Weighting is done by inverse document frequency (lowers the impact of frequently occurring words.
  - Words occurring in less than 15% of the documents are excluded.
  - All words except "light" that occur in over 85% of the texts are excluded.
  - FYI, we have also used singular value decomposition and a Bayesian Dirichlet classifier but they are more difficult to interpret and are not presented here.
- The Bayesian clustering identifies 6 clusters.

Cluster Plot of EECBG Grant Text - 6 Clusters



These two components explain 76.19 % of the point variability.

Identifying Cluster Meaning

- We use words particular to each cluster to determine meaning.
  1. Residential and business power efficiency (efficiency, home, residential, commercial, contractor, power usage)
  2. Audit (audit, reduce waste, inform, window)
  3. Solar (Solar, power, generate, house)
  4. Retrofit (retrofit, conserve, heat, construct)
  5. Economic Development (job, fuel, growth)
  6. Management (budget, monitor, resource, no “tribal”)

Using Clustering to Investigate Economic Development

- After retrieving the clusters, we can include them in our regression analysis.

- If we use the respondent-defined categorization of the grant as clustering variables (fixed effects) we get:
  - A large amount of unnecessary variation in the model from inconsistent application.
  - Insignificant impact for economic development.

- After we switch to text mining developed clusters, economic development does become more precise.

Caveats

- Before I demonstrate the preliminary regression results:

1. The clusters were calculated separately than the rest of the data set and then merged up using fuzzy matching.
    - No unique identifier to merge records. Grant number repeats and may have several different texts associated with it)
    - There were software license difficulties that were not resolved early enough to correct this. I apologize.

2. The text analysis results are not "publication ready"
    - Should use a training set and a final set, which I plan on doing once I can properly merge the data before beginning.
    - I plan on using the smaller grants as a training set, then the larger grants in the final analysis.

Results

	Original Measurement		Text Clusters	
Variables	Estimate	SE	Estimate	SE
Satisfaction w/DOE Application Process	-0.079	0.024	-0.087	0.026
Satisfaction w/DOE Approval Process	0.021	0.033	0.028	0.037
Satisfaction w/DOE Tech. Support	-0.139	0.036	-0.127	0.040
Administrative Capacity	-0.016	0.063	-0.011	0.059
External application assistance	0.034	0.033	0.034	0.034
Citizen application participation	-0.098	0.057	-0.105	0.075
Copied policies from other governments	0.096	0.046	0.091	0.067
Innovative (new) policies to implement	0.081	0.141	0.086	0.170
Citizen advocacy level	-0.009	0.067	-0.011	0.085
Number of prior sustainable policies	-0.042	0.033	-0.033	0.037
Green practices count	0.071	0.042	0.058	0.023
Green development in planning	-0.054	0.053	-0.069	0.089
Economic development tool	-0.063	0.038	-0.071	0.029
Budget (logged)	0.003	0.013	0.003	0.017
Unemployment	0.018	0.014	0.016	0.023
Manager Form of Government	-0.236	0.235	-0.238	0.201

Fixed effects DOE				
Clean energy policy	0.463	0.180		
Financial incentives for energy efficiency and other covered investments	0.095	0.236		
Government, school, institutional procurement	-0.364	0.241		
Loans and grants	-0.010	0.148		
Renewable energy market development	-0.137	0.233		
Workshops, training, education	0.191	0.183		
Building energy audits	0.345	0.221		
Other	0.243	0.114		
Technical Assistance	0.459	0.145		
Transportation	-0.178	0.122		
Fixed Effects Cluster				
1.Residential and business power efficiency			-0.174	0.072
2.Audit			0.178	0.051
3.Solar			0.257	0.104
4.Retrofit			0.395	0.200
5.Economic Development			0.318	0.112
6.Management			-0.256	0.293

Discussion

- In addition to statistical significance (which is meaningless!), the model fit statistics (AIC, SBIC) clearly show the models with text based clustering increases the amount of variation explained.
- We believe these results demonstrate that text analysis could be used to better group and control for random variation in the model.
- The clusters predicted by text analysis provide a better sense of which projects are similar so that modeling can accurately account for heterogeneity.

Expanding the Methodology

- The EECBG grants are futile training grounds for expanding this methodology into other projects.
  1. The publication we are working on using these methods will actually look at the interaction of economic development motivation with the different classes of grants.
  2. Use key words as a training data set to analyze other sustainability grant text.
     - Likely at the state level
  3. Fuel for some qualitative analysis of outliers (Jessica).
  4. Look at grant changes within the EECBG program.
     - About 3% of the grants changed in this program.
     - Look to see if they adopted language similar to original submissions.