Data Driven Mathematical Modeling

Rodney X. Sturdivant and Robert E. Burks

JMM 2020

The "Data Insights" Problem (MCM Problem C)

Problem C Overview

- Started in 2016
- "Amplify" modeling challenges associated with data
- Not necessarily "big data"
- Complicating factors
  - Size, data types, missing etc.
- Includes data files

Problems

- *2016 "The Goodgrant Challenge"*
- *2017 "Cooperate and Navigate"*
- *2018 "Energy Production"*
- *2019 "The Opioid Crisis"*

Meet the data

2016 “The Goodgrant Challenge”

- $100 million grant money
- Donate to a group of schools over 5 years
- Optimal allocation to improve “ROI”
  - Goal: student performance improvement
- Produce a prioritized list of schools for each year

DATA

- U.S. National Center on Education Statistics
  - Survey data
- College Scorecard
  - Performance data
- 122 data elements
- 7800+ schools

Meet the data

2017 “Cooperate and Navigate”

- Effects of introducing self-driving cars
- State of Washington
  - I-5, I-90, I-405, SR520
- Model effects, propose policies
- Dedicated lanes, percentage of self-driving cars, peak vs normal hours, interactions

DATA

- 4 roads
- Average cars per day driving on road
- Data available for each milepost on the road
  - 224 mileposts
- Number of lanes (at each milepost)
  - Number of lanes in “increasing direction”

Meet the data

2018 “Energy Production”

- 4 states: AZ, CA, NM, TX
- Develop energy profile for each state, model the profile over time
- Determine the “best profile” (renewable energy)
- Predictions and targets for each state – actions to meet goals

DATA

- 50 years
- 605 variables
- Energy production and consumption

Meet the data

2019 "The Opioid Crisis"

- Spread and characteristics of synthetic opioids/heroin
  - Patterns, concerns, thresholds, origins
- Socioeconomic factors
- Develop and test strategy
- 5 states (OH, PA, KY, VA, TN)

DATA

- County level data – 462 counties
- 2010 – 2017
- 69 drugs, total cases
- Socio economic data (census) by year
  - 150+ variables
- Not provided (but allowed)
  - Map data: coordinates/distances

Data Specific Challenges

- Exploratory Data Analysis (EDA)

"Data Wrangling"

  - Data processing
  - Data cleaning
  - Data visualization

Data Processing and Cleaning

- Handling data types
  - Text – State
  - Numerical two categories (binary) – HBCU
  - Numerical many categories – LOCALE (large city, fringe suburb etc)
  - Numerical and continuous/discrete – SATV, PCIP

Goodgrant Challenge Example

STABBR	NPCURL	HCM2	PREDDEG	LOCALE	HBCU	SATVR25	PCIP11	GRAD_DEBT_MDN_SUPP	ACTCM25
AL	galileo.aamu.edu/netpricecalculator/np	0	3	12	1	370	0.0348	33611.5	15
AL	www.collegeportraits.org/AL/UAB/esti	0	3	12	0	520	0.0099	23117	22
AL	tcc.noellevitz.com/(S(miwoihs5stz5cpyi	0	3	12	0	NULL	0.0411	PrivacySuppressed	NULL
AL	finaid.uah.edu/	0	3	12	0	510	0.0273	24738	23

- Scales (PCIP – percentage of degrees in computer science vs Grad Debt)
  - Standardization may be required
- Different indicators of missing (some years many!)
  - Some “missing” data actually tells us something

RELAFFIL	-1	Not reported
	-2	Not applicable
	22	American Evangelical Lutheran Church
	24	African Methodist Episcopal Zion Church
	27	Assemblies of God Church

Data Processing and Cleaning

- Outliers and unusual observations
  - Identification (visualizations etc)
  - Could be legitimate and important data…
  - Could be legitimate but not important (ex: not a value we care to predict)…
  - Could be bad data

“Goodgrant Challenge” Example

Schools with **negative** average tuition

Opiod Crisis Example

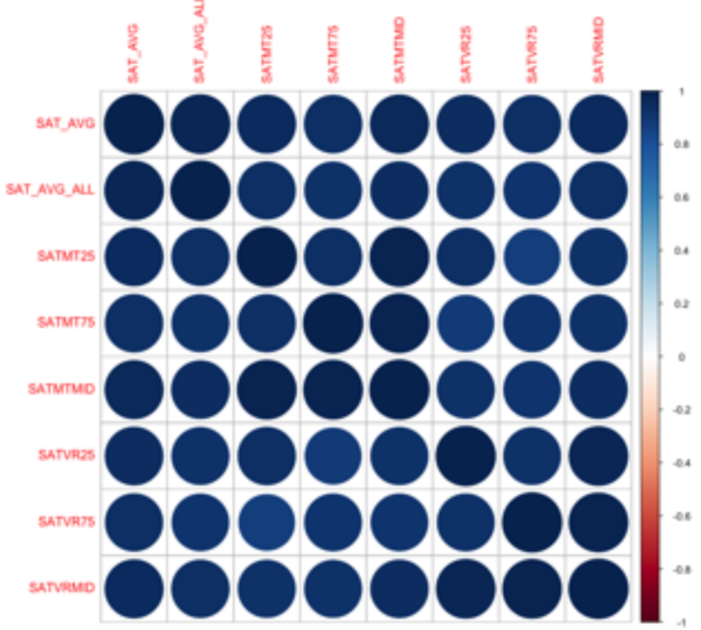
Shift in census designations in 2013

Great “catches”
Most teams did not notice these issues!

Advanced Data Processing and Cleaning

- Highly correlated data
  - May cause issues in some models (multicollinearity)
  - ”Redundant”
  - Some teams handled with PCA or other dimension reduction approach (not always intentionally)

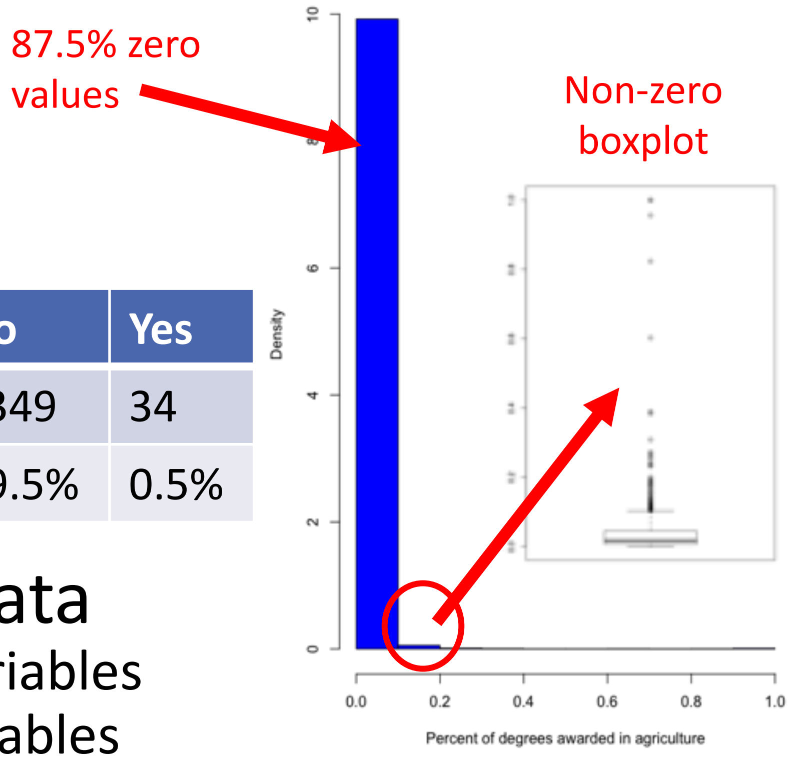
SAT scores – Pearson correlations > 0.9



“Goodgrant Challenge” Examples

TRIBAL	No	Yes
Count	7349	34
Percent	99.5%	0.5%

Percent of degrees awarded in agriculture



- Highly skewed data
  - “Zero-variance” variables
  - “Zero inflated” variables
  - Often hurts model building, model assumptions
  - Very few teams recognize and intentionally address

Dimension Reduction Methods

- **Principal Components Analysis (PCA)**
  - Create a smaller set of ”components” that accounts for a high percentage of variability in data
  - Components are linear combinations of original variables
  - Eigenvectors corresponding to largest eigenvalues

Goodgrant Challenge: school performance measures
Opioid Crisis: socioeconomic factors

- Identification of PC’s (how many to use)
- Interpretation (not always easy, not always done properly!)

- **Cluster Analysis**
  - n observations each a d-dimensional vector of values
  - Choose k sets of observations (clusters) to minimize the distances of points from the center of their cluster

Goodgrant Challenge: groups of similar schools
Opioid Crisis: groups of counties

- Choice of k
- Are observations in clusters really similar

Handling Missing Data

- Exclude variable or observation
- Simple imputation
  - Example: missing drug count data (opioid data) impute a 0
  - Replace with the an average (mean, median) value
- Use of regression models
  - Built with other variables as predictors
- Multiple Imputation (MI)

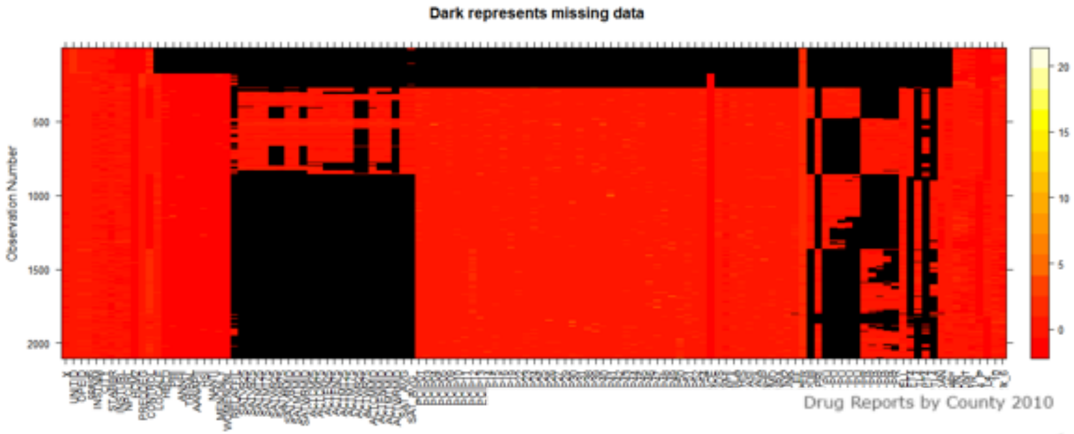
1. Missing data are filled in *n* times, generating *n* complete data sets.
2. Each complete data set is analyzed using a statistical procedure.
3. The results of the analysis from each of the *n* complete data sets are combined for the statistical inference.

- Decision on how to treat matters (example: 0 means something)
- Missingness mechanism matters
  - MCAR (Missing Completely at Random) – simple may be ok
  - MAR (Missing at Random) – missingness related to observed data – use regression or MI
  - MNAR (Missing Not at Random) – missing data itself related to why missing – danger in imputation

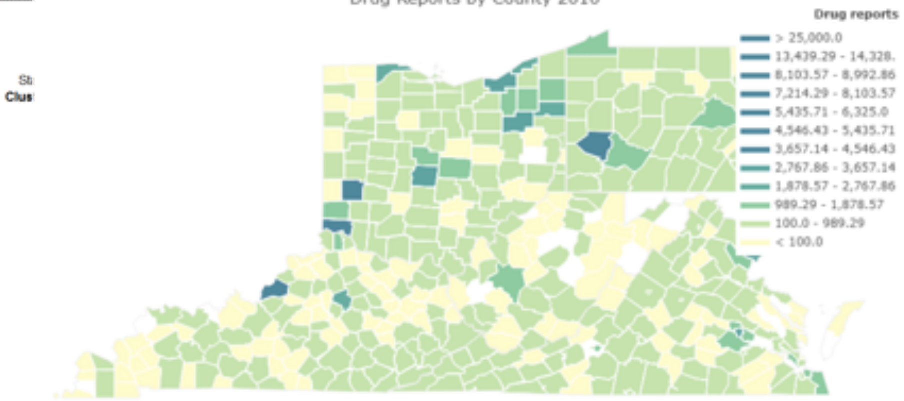
Data Visualization

- Standard methods: trendlines, scatterplots, histograms, boxplots, etc.
- Many very cool plots (radial, heat map, etc)
- Useful at all stages of the modeling!

EXPLORING DATA

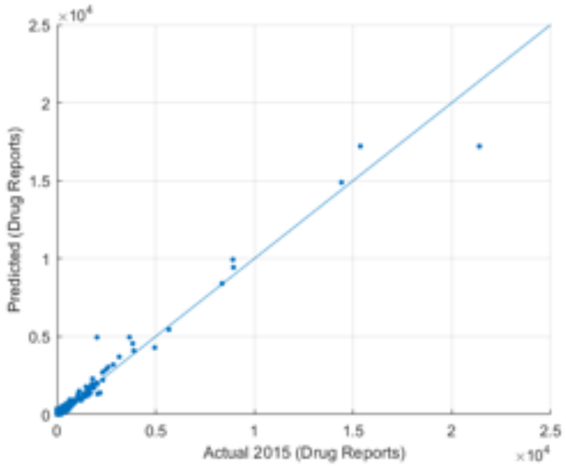


Find missing data issues



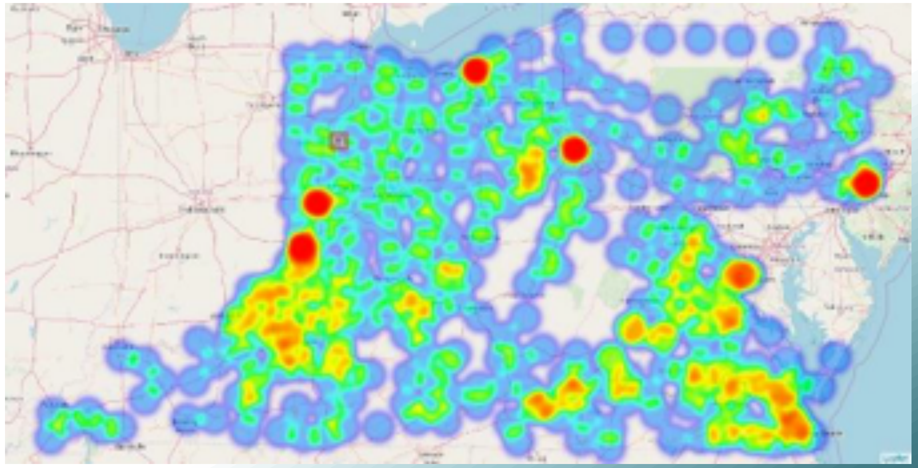
Raw data (drug reports)

PRESENTING RESULTS



Actual vs predicted

Model results of opioid spread



MODELS

A VERY wide variety in most years! A few examples…

2016 "The Goodgrant Challenge"

- Analytic Hierarchy Process (AHP)
- LASSO
- Bayes
- PCA
- Linear regression
- Cluster analysis

Usually more than one technique
- Submodels
- Parts of problem

2017 "Cooperate and Navigate"

- Cellular Automata
- Fluid flow (differential equations)

Less variety this year – CA by far most common!

2018 "Energy Production"

- Neural Nets
- Entropy
- ARIMA
- TOPSIS
- Correlation/regression
- AHP/PCA

HUGE variety these years – others too!

2019 "The Opioid Crisis"

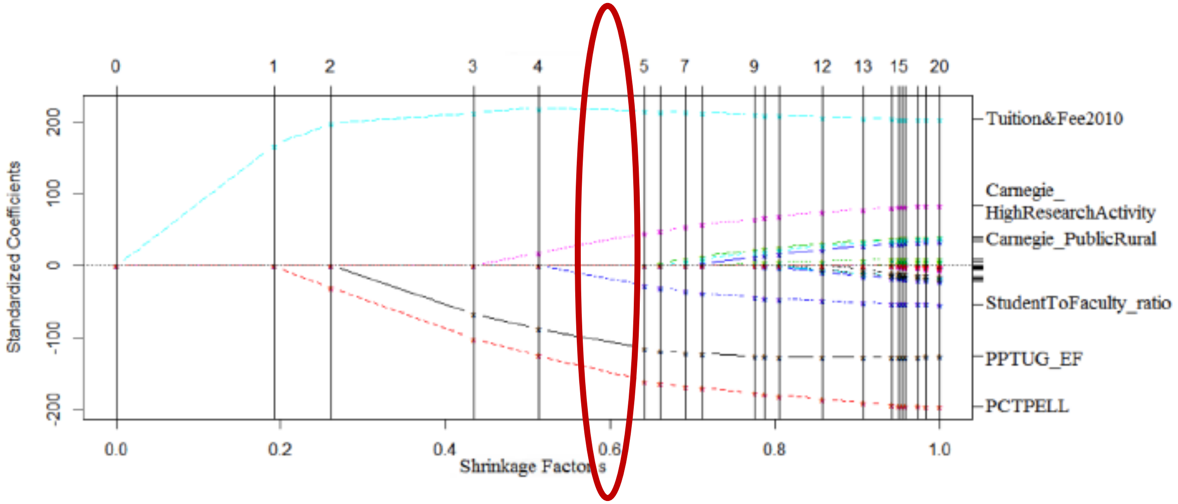
- Time series models (Gray, ARIMA)
- SIR (differential/difference equations)
- Markov simulations
- Support Vector Machines (SVM)
- Regression/ANOVA
- CART and Random Forests

LASSO (least absolute shrinkage and selection operator)

”Goodgrant Challenge” Example*

* Tsinghua University, China. Title: An Optimal Strategy of Donation for Educational Purpose. 2016 MCM entry.

- A “performance index” (PI) developed to measure school effectiveness
  - Use PCA – a weighted average of several metrics
- Want a model to determine which of 108 indicators best predict the PI



“Penalized” Linear Regression Model

- Coefficients of predictors minimize:

$$\min_{\beta} \frac{1}{n}\sum_{i}(y_i - x_i^T \beta_i)^2 + \lambda \| \beta \|.$$

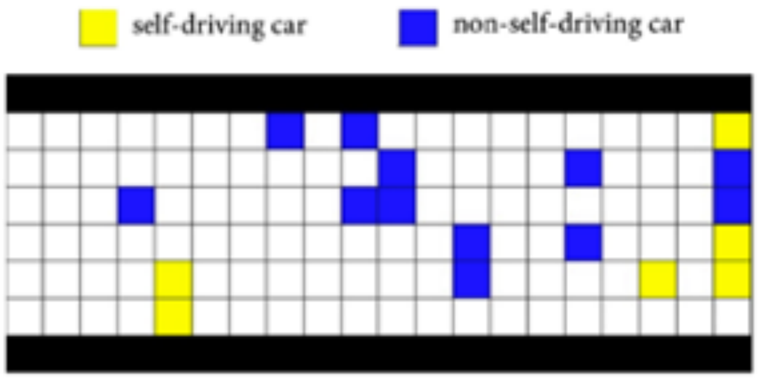
- Idea: avoid overfitting by adding penalty for large coefficients
- Shrinkage parameter, λ, selected to minimize Mean Square Error (MSE) for predictions in cross validation
  - 10—fold CV: hold out a different 10% of sample each time
  - Use model based on 90% to predict the hold out sample

- Determined an optimal shrinkage at λ = 0.6
- 5 predictors with non-zero coefficients
  - Example negative coefficient for student to faculty ratio – makes sense!

Cellular Automata (CA)

- Discrete time simulation
- Grid of simple elements (cells)
  - Assume one of two states
- At each step retain current state or transition based on a set of rules
- Rules based on cell and neighboring cell information
- ”Cooperate and Navigate” most used
  - also popular in ”Opiod Crisis”

”Cooperate and Navigate” Example

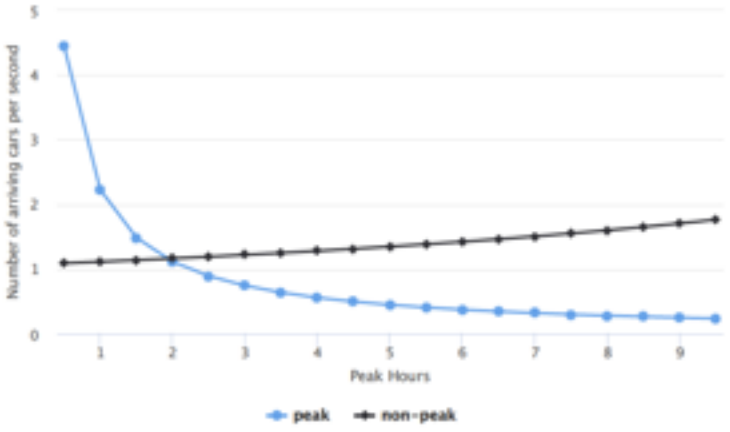


Cell definitions/attributes

- Lengths of road/lane
- Occupied or not
- Type of car (SD or not)
- Car characteristics at time: velocity, acceleration, turn signal etc.

Vehicle generation

- Constant arrivals
- Using data provided (few teams did this!)
- Poisson arrivals
- Example shown: bimodal Gaussian (peak and non-peak arrival rates - resulting probabilities for number cars per second)



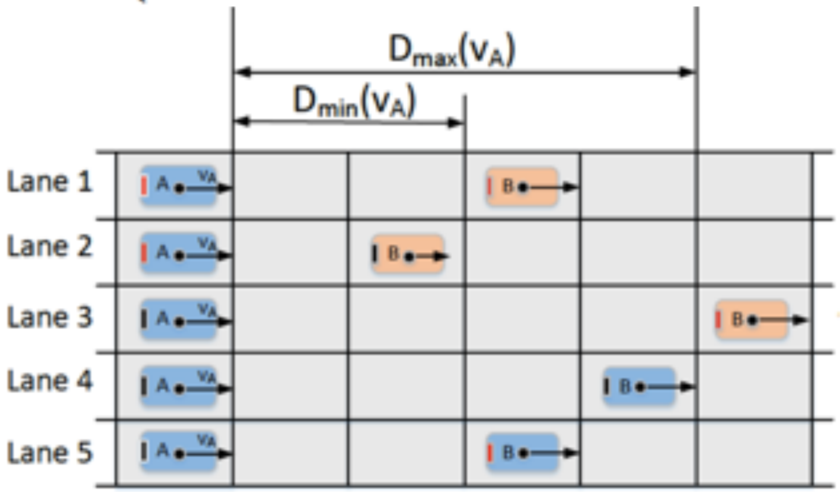
Cellular Automata (CA)

”Cooperate and Navigate” Examples of Rules

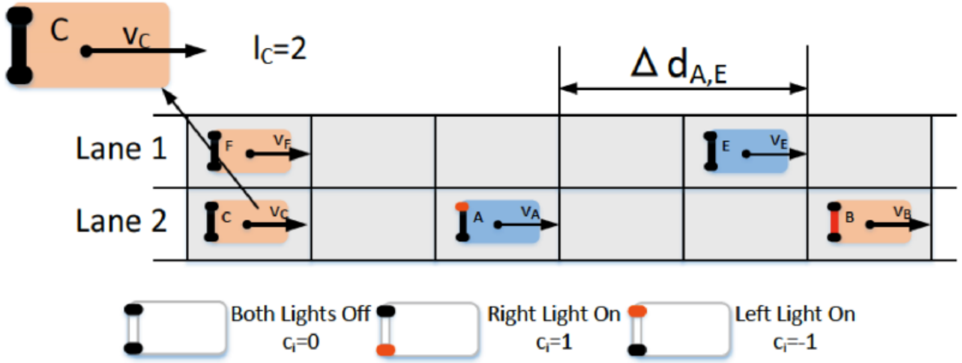
“Following Rules”

- Accelerate/decelerate probabilities
- Distance of car ahead
- Other safety indicators
  - Ex: brake lights

$$P_{decelerate} = \begin{cases} 0.94 & \text{B brake lights on and } D_{safe} < d < D_{max} \\ 0.5 & \text{B brake lights off and } D_{safe} < d < D_{max} \\ 0.2 & v = 0 \\ 1 & d < D_{safe} \\ 0 & \text{otherwise} \end{cases}$$



“Lane Change Rules”



“Highway Interchange Rules”

- Very few teams attempted

SELF DRIVING CARS different parameters and rules

CART (Classification and Regression Trees)

"Opioid Crisis" Example*

Predictors (socioeconomic factors) of total opioid reports

1. Determine each factor's best "split" for the data in a given "leaf" (node) of "tree".
2. Pick the factor giving the best split.
3. Split the data in the given leaf/node based on the chosen factor and split point.
    - If no stopping rules are met
4. Use the mean drug reports within each created leaf as the predicted value for counties in that leaf.

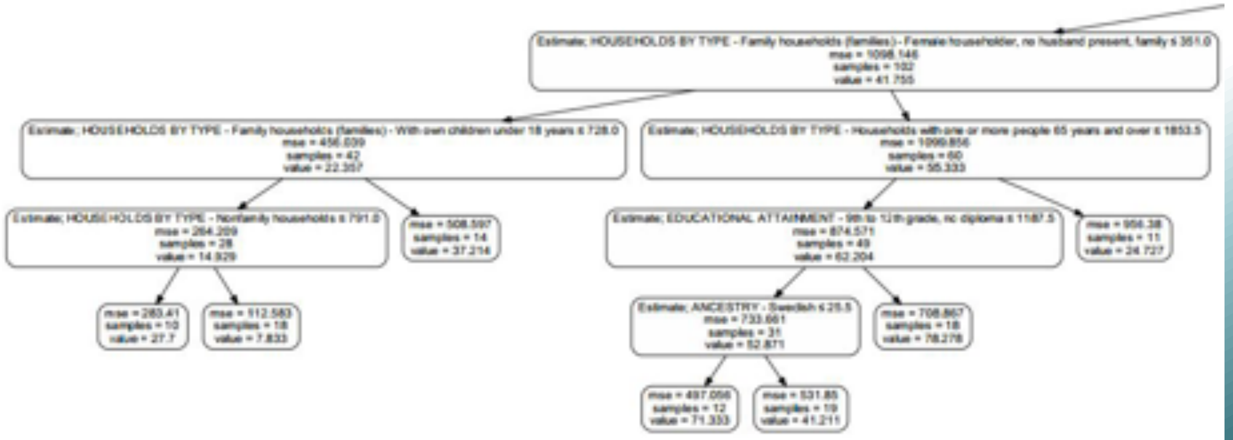
Iterate until no further splits possible

Factors in tree

MARITAL STATUS
EDUCATIONAL ATTAINMENT
HOUSEHOLDS BY TYPE
DISABILITY STATUS
SCHOOL ENROLLMENT
PLACE OF BIRTH
GRANDPARENTS
RELATIONSHIPS IN HOUSEHOI
RESIDENCE 1 YEAR AGO
VETERAN STATUS
FERTILITY
YEAR OF ENTRY

CART considerations

- Criteria to measure predictive accuracy/determine optimal splits
    - Team used Mean Squared Error (MSE)
- Tuning parameters related to tree "depth"
    - Number of observations in leaf or depth
- OVERFITTING concern – "prune" tree



* Sichuan University, China. Title: Opioid Use Profile and Recommended Strategies in 5 States. 2019 MCM entry.

RANDOM FOREST

"Opioid Crisis" Example*

Build *m* trees and average for prediction

To build each tree:

1. Generate a bootstrap sample of original data
2. Create a tree for this sample
   - For each split randomly select *k* predictors
   - Select best of the *k* to make split
3. Build complete tree (without pruning) with typical stopping criteria

RF considerations

- Tuning parameters number of trees (*m*) and predictors for possible splits (*k*)
  - Various measures and methods to optimally choose
- Idea is to create uncorrelated trees
- OVERFITTING not a concern
- Methods to identify most important predictors
  - Exact nature of relationships not clear – hard to interpret model

TOP 10
most important socioeconomic factors using RF

total illicit drug use rate	total population
people born in the US	American ancestry
Irish ancestry	only English spoken at home
some college but no degree	high school graduation rate
Polish ancestry	graduate or professional degree.

90% accuracy predicting test data

* University of Colorado Boulder, CO, USA. RandomWalks and Rehab: Analyzing the Spread of the Opioid Crisis. 2019 MCM entry.

Sensitivity analysis and model validation

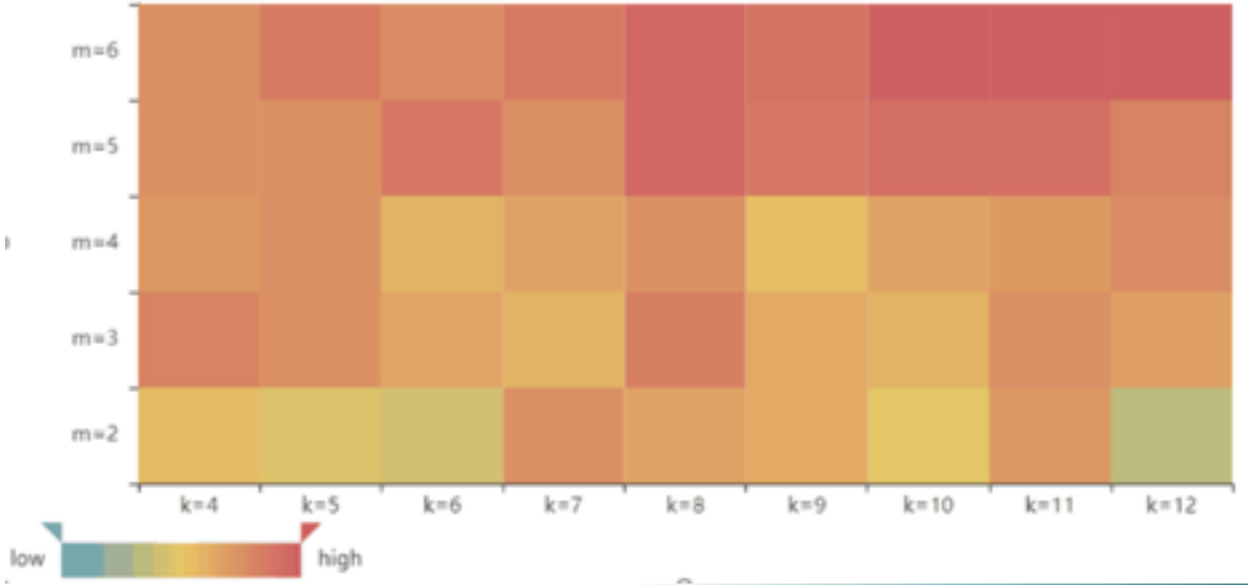
Approaches may differ from typical MCM problems

- Tests/measures of uncertainty
  - Confidence intervals, statistical tests
- Measures of model predictive performance
  - R^2 , RMSE, sensitivity/specificity, ROC curves
- Statistical models goodness-of-fit
  - Residual analysis
- Methods of finding optimal tuning parameters
  - Cross-validation

- Simulation methods (CA) typically use sensitivity analysis of key parameters

- Generally, should involve model and DATA
  - Often not done or done poorly!

Example: ”Opioid Crisis” CA model for spread*
- 2 parameters (k - # of counties considered “neighbors”, m – “environment” of county)
- Examine RMSE of predicting opioid counts



* University of Electronic Science and Technology of China, China. The Current Status, Future and Strategy of Opioid. 2019 MCM entry.

"Statisticians are working thru the night processing numbers after an explosion of data in biotechnology has trapped a dozen data miners."

QUESTIONS?