Data, Environment and Society: Lecture 16: Classification

Instructor: Duncan Callaway
GSI: Salma Elmallah

October 24, 2019

Upcoming lectures

Today

- Introduction to classification (this slide deck)
  - Corresponding reading: ISLR Ch 4.1 through 4.3.1, and also Ch 2.2.3.
- Intro to classification and regression trees (other slide deck)

Next week

- Tuesday:
  - Dan Kammen: Racial dispartities in rooftop solar PV adoption
    - Reading: Sunter et al
  - Salma: Intro to Environmental Justice
- Thursday:
  - Duncan: Finish discussing EJ
    - Reading: Pastor
  - Finish classification and regression trees
    - Reading: ISLR 8.1-8.2

Remember qualitative variables? This is how you do it

$$x_1 = \begin{cases} 1, & \text{Likes split pea.} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_2 = \begin{cases} 1, & \text{Likes minestrone.} \\ 0, & \text{otherwise.} \end{cases}$$

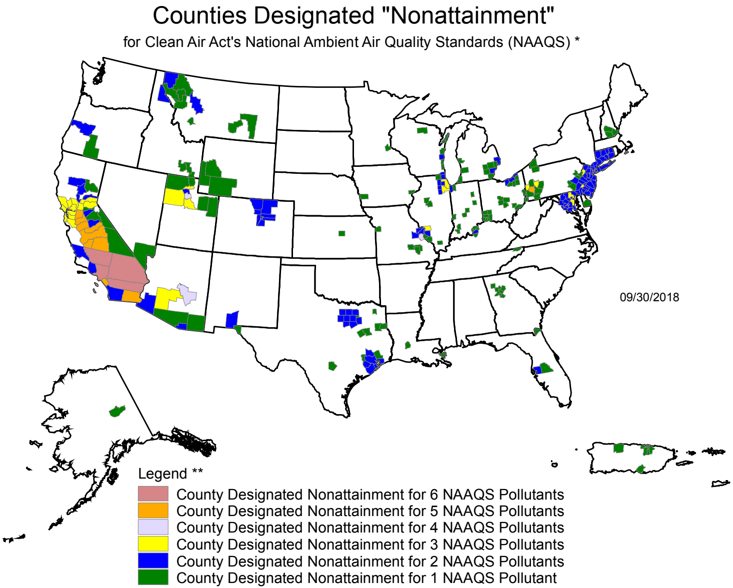
$$x_3 = \begin{cases} 1, & \text{Doesn't like soup.} \\ 0, & \text{otherwise.} \end{cases}$$

Question: What about the "other" category?
Answer: The answers are mutually exclusive, so if x_1, x_2, x_3 are all zero, then the answer must be "other".

What it you want to *predict* a categorical variable?

Examples:

- Clean air act attainment status for a region (attainment or non-attainment)
- Is an area going to be a candidate for a new refinery in the next 10 years
- Disease presented in emergency dept of a hospital



Why not linear regression?

For example: EPA Criteria Pollutants. For each region i indicate nonattainment as follows:

- $y_i = 1 \rightarrow$ Ozone
- $y_i = 2 \rightarrow$ PM2.5
- $y_i = 3 \rightarrow$ PM10
- $y_i = 4 \rightarrow$ Sulfur Dioxide
- $y_i = 5 \rightarrow$ Lead
- $y_i = 6 \rightarrow$ Carbon Monoxide
- $y_i = 7 \rightarrow$ Nitrogen Dioxide

Then,

$$y_i = \beta \mathbf{x}_i + \epsilon$$

where $\mathbf{x}_i$ is a vector of observed independent variables for each location i

Why not linear regression?

For example: EPA Criteria Pollutants. For each region i indicate nonattainment as follows:

- $y_i = 1 \rightarrow$ Ozone
- $y_i = 2 \rightarrow$ PM2.5
- $y_i = 3 \rightarrow$ PM10
- $y_i = 4 \rightarrow$ Sulfur Dioxide
- $y_i = 5 \rightarrow$ Lead
- $y_i = 6 \rightarrow$ Carbon Monoxide
- $y_i = 7 \rightarrow$ Nitrogen Dioxide

Then,

$$y_i = \beta \mathbf{x}_i + \epsilon$$

where $\mathbf{x}_i$ is a vector of observed independent variables for each location i

The problem: a different ordering of variables would imply a different relationship between statuses, different models, and different predictions.

Predicting categorical variables the right way: Similar coding to using them as predictors

For example: EPA Criteria Pollutants:

- $y_{i,1} \rightarrow$ Ozone status
- $y_{i,2} \rightarrow$ PM2.5 status
- $y_{i,3} \rightarrow$ PM10 status
- $y_{i,4} \rightarrow$ Sulfur Dioxide status
- $y_{i,5} \rightarrow$ Lead status
- $y_{i,6} \rightarrow$ Carbon Monoxide status
- $y_{i,7} \rightarrow$ Nitrogen Dioxide status

where *i* indexes the region of interest, i.e. observations

Then, for each,

$$y_{i,j} = \begin{cases} 1, & \text{Nonattainment status} \\ 0, & \text{Attainment status} \end{cases}$$

where *j* indexes criteria pollutants

How do you get a model to output a $\{0, 1\}$ result?

How do you get a model to output a $\{0, 1\}$ result?

First, build your model so that its output estimates a *probability* that a given outcome happens.

For example, a model might give:

$p(\text{PM-2.5 nonattainment}) = 0.734$ and $p(\text{PM-2.5 attainment}) = 0.266$

(the probabilities add to one).

How do you get a model to output a $\{0, 1\}$ result? (ctd)

Second, we use something called the *Bayes Classifier*.

Bayes classifier:

For a given observation x_i, choose j as the value for which

$$\Pr(Y = j | X = x_i)$$

"conditional on"

is largest.

More formally,

How do you get a model to output a $\{0, 1\}$ result? (ctd)

Second, we use something called the *Bayes Classifier*.

Bayes classifier:

For a given observation x_i, choose j as the value for which

$$\Pr(Y = j | X = x_i)$$

is largest.

More formally,

observed features

argument

$$\hat{y}_i = \arg\max_{j \in \mathcal{J}} \underline{\Pr(Y = j | X = x_i)}$$

where $\mathcal{J}$ is the set of possible (mutually exclusive) outcomes.

Classification error rate

If the "true model" for $\Pr(Y = j | X = x_i)$ is known, then using the Bayes classifier will provide the lowest possible error rate.

The *error rate* quantifies how frequently a model mis-classifies a categorical variable.

Let $I(\cdot)$ denote the *indicator function*:

$$I(y_i \neq \hat{y}_i) = \begin{cases} 1, & \text{if } y_i \neq \hat{y}_i \\ 0, & \text{otherwise} \end{cases}$$

Classification error rate

If the "true model" for $\Pr(Y = j | X = x_i)$ is known, then using the Bayes classifier will provide the lowest possible error rate.

The *error rate* quantifies how frequently a model mis-classifies a categorical variable.

Let $I(\cdot)$ denote the *indicator function*:

$$I(y_i \neq \hat{y}_i) = \begin{cases} 1, & \text{when } y_i \neq \hat{y}_i. \\ 0, & \text{otherwise.} \end{cases}$$

$$\Rightarrow \text{error rate} = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

What's the error rate?

Location	Actual status	Predicted (hypothetical)	$I(y_i \neq \hat{y}_i)$
Allegheny County, PA	Non-attainment	Non-attainment	
Cleveland, OH	Non-attainment	Non-attainment	
Delaware County, PA	Non-attainment	Non-attainment	
Imperial County, CA	Non-attainment	Attainment	
Lebanon County, PA	Non-attainment	Attainment	
South Coast Air Basin, CA	Non-attainment	Non-attainment	
Plumas County, CA	Non-attainment	Non-attainment	
Sacramento County, CA	Attainment	Non-attainment	
San Francisco, CA	Attainment	Non-attainment	
San Joaquin Valley, CA	Non-attainment	Non-attainment	
West Silver Valley, ID	Non-attainment	Non-attainment	
Luzerne County, PA	Attainment	Attainment	
Lancaster County, PA	Attainment	Attainment	

What's the error rate?

Location	Actual status	Predicted (hypothetical)	$I(y_i \neq \hat{y}_i)$
Allegheny County, PA	Non-attainment	Non-attainment	0
Cleveland, OH	Non-attainment	Non-attainment	0
Delaware County, PA	Non-attainment	Non-attainment	0
Imperial County, CA	Non-attainment	Attainment	
Lebanon County, PA	Non-attainment	Attainment	
South Coast Air Basin, CA	Non-attainment	Non-attainment	
Plumas County, CA	Non-attainment	Non-attainment	
Sacramento County, CA	Attainment	Non-attainment	
San Francisco, CA	Attainment	Non-attainment	
San Joaquin Valley, CA	Non-attainment	Non-attainment	
West Silver Valley, ID	Non-attainment	Non-attainment	
Luzerne County, PA	Attainment	Attainment	
Lancaster County, PA	Attainment	Attainment	

What's the error rate?

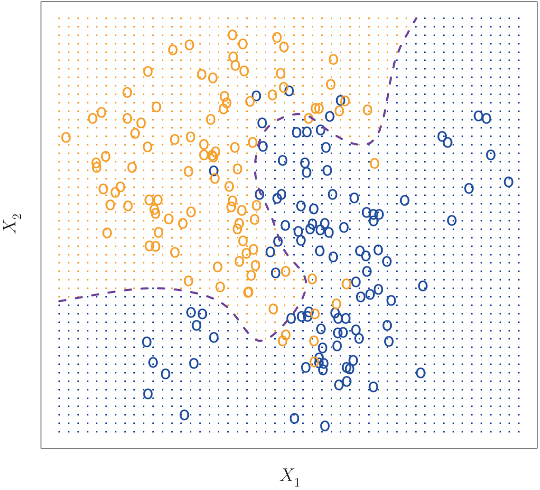
Location	Actual status	Predicted (hypothetical)	$I(y_i \neq \hat{y}_i)$
Allegheny County, PA	Non-attainment	Non-attainment	0
Cleveland, OH	Non-attainment	Non-attainment	0
Delaware County, PA	Non-attainment	Non-attainment	0
Imperial County, CA	Non-attainment	Attainment	1
Lebanon County, PA	Non-attainment	Attainment	1
South Coast Air Basin, CA	Non-attainment	Non-attainment	0
Plumas County, CA	Non-attainment	Non-attainment	0
Sacramento County, CA	Attainment	Non-attainment	
San Francisco, CA	Attainment	Non-attainment	
San Joaquin Valley, CA	Non-attainment	Non-attainment	
West Silver Valley, ID	Non-attainment	Non-attainment	
Luzerne County, PA	Attainment	Attainment	
Lancaster County, PA	Attainment	Attainment	

What's the error rate?

Location	**Actual status**	**Predicted (hypothetical)**	$I(y_i \neq \hat{y}_i)$
Allegheny County, PA	Non-attainment	Non-attainment	0
Cleveland, OH	Non-attainment	Non-attainment	0
Delaware County, PA	Non-attainment	Non-attainment	0
Imperial County, CA	Non-attainment	Attainment	1
Lebanon County, PA	Non-attainment	Attainment	1
South Coast Air Basin, CA	Non-attainment	Non-attainment	0
Plumas County, CA	Non-attainment	Non-attainment	0
Sacramento County, CA	Attainment	Non-attainment	1
San Francisco, CA	Attainment	Non-attainment	1
San Joaquin Valley, CA	Non-attainment	Non-attainment	0
West Silver Valley, ID	Non-attainment	Non-attainment	0
Luzerne County, PA	Attainment	Attainment	0
Lancaster County, PA	Attainment	Attainment	0

$\rightarrow$ Overall error rate is $\frac{1}{13} \cdot 4 = .31$

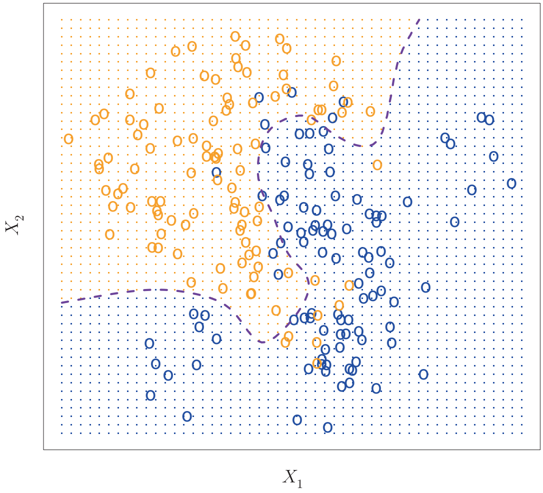
Finally, the decision boundary



ISLR Fig 2.13

Along the boundary, what condition holds?

Finally, the decision boundary



ISLR Fig 2.13

Along the boundary, what condition holds? **The probability of one outcome equals the probability of the other.**

If the true model $\Pr(Y = j|X = x_i)$ is known, we call this the **Bayes decision boundary.**

All methods try to *estimate* the Bayes decision boundary.

KNN for categorical variables

It's pretty simple! Estimate the probabilities as

$$\Pr(Y = j \mid X = x_0) = \frac{1}{K} \sum_{k \in \mathcal{N}_0} I(Y_k = j)$$

\# neighbors

index neighbor

K = number of neighboring ***training*** points to consider
$\mathcal{N}_0$ = set of K ***training*** points closest to observation x_0.

...Then apply the Bayes classifier to choose the outcome variable.

KNN for categorical variables

It's pretty simple! Estimate the probabilities as

$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

$K =$ number of neighboring ***training*** points to consider
$\mathcal{N}_0 =$ set of K ***training*** points closest to observation x_0.

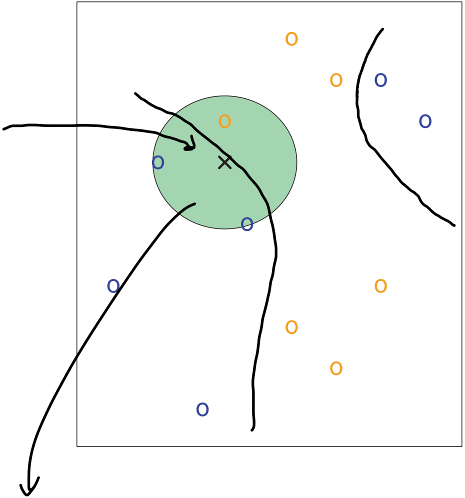
...Then apply the Bayes classifier to choose the outcome variable.

$$\hat{y}_i = \arg\max_{j \in \mathcal{J}} \Pr(Y = j|X = x_i)$$

where x_i is any feasible point in the space of independent variables.

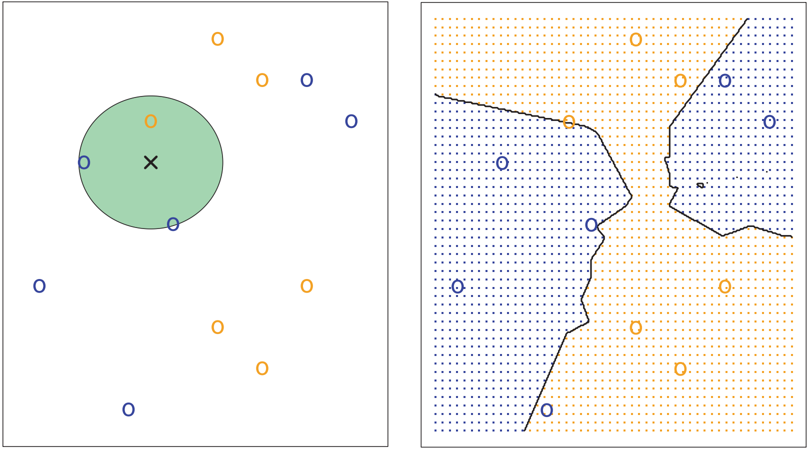
Simple KNN Example, $K = 3$

x_0



$P(Y = \text{blue} \mid X = x_0) = \frac{2}{3}$ $\quad$ $P(Y = \text{gold} \mid X = x_0) = \frac{1}{3}$ $\quad \Rightarrow \quad \hat{Y}_0 = \text{blue}$

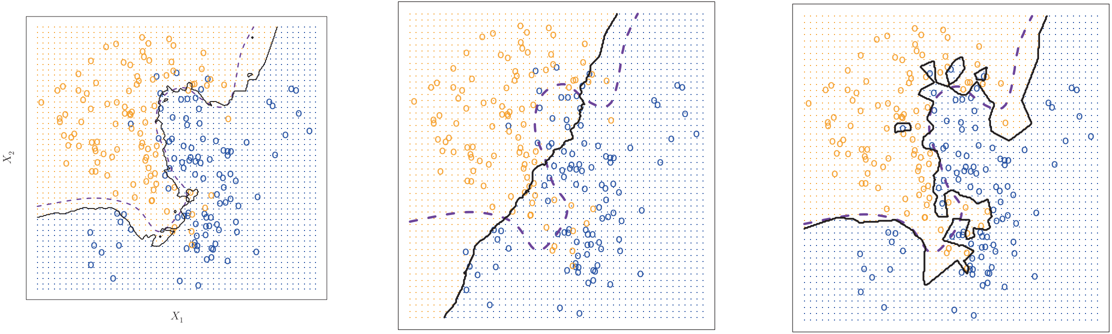
Simple KNN Example, $K = 3$



ISLR Fig 2.14

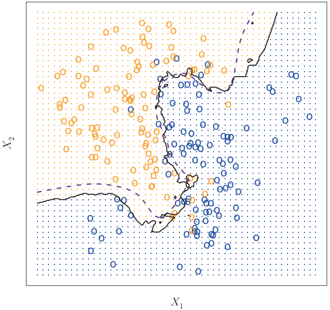
Which has the highest K? Which has the lowest?

Dashed = Bayes decision boundary
Solid = KNN estimate of Bayes decision boundary

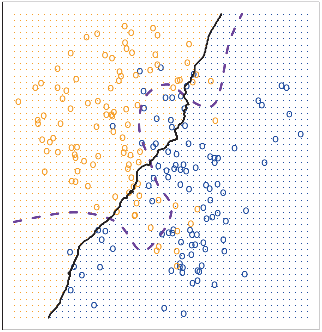


Which has the highest K? Which has the lowest?

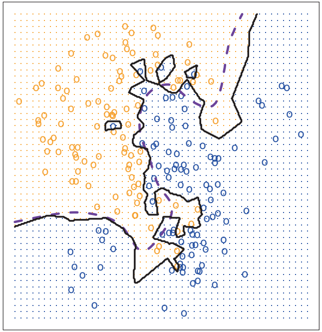
Dashed = Bayes decision boundary
Solid = KNN estimate of Bayes decision boundary



K = 10

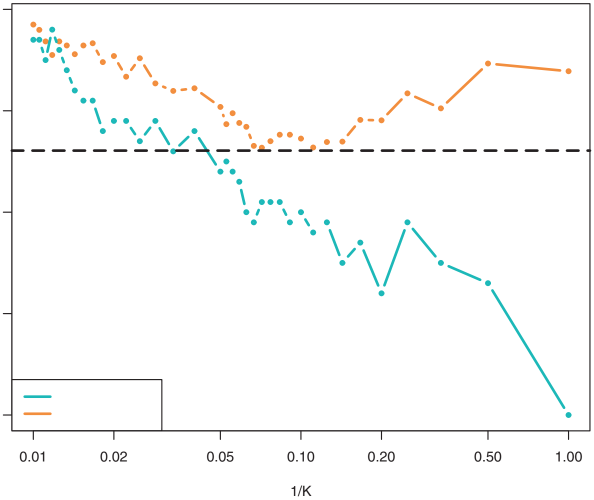


K = 100



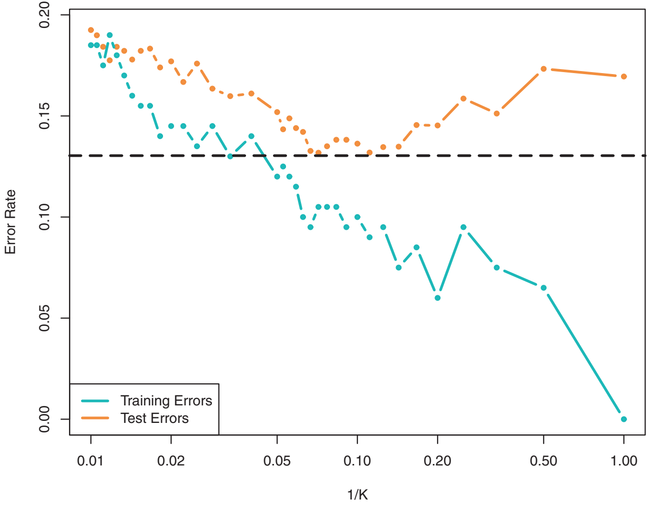
K = 1

KNN test and training error



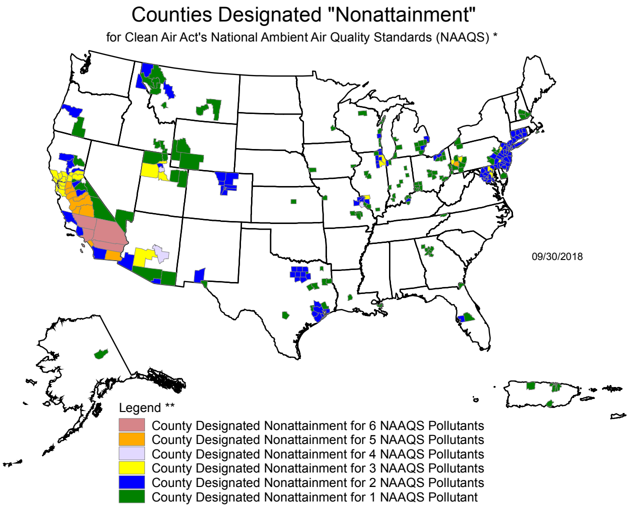
ISLR 2.17

KNN test and training error

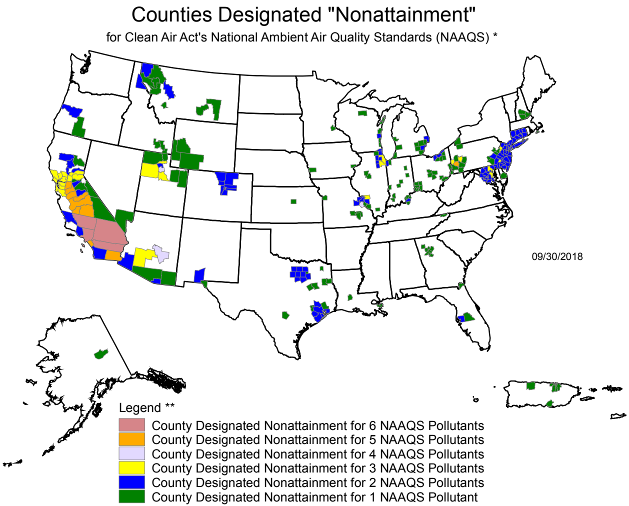


ISLR 2.17

How would KNN perform with nonattainment areas?



How would KNN perform with nonattainment areas?



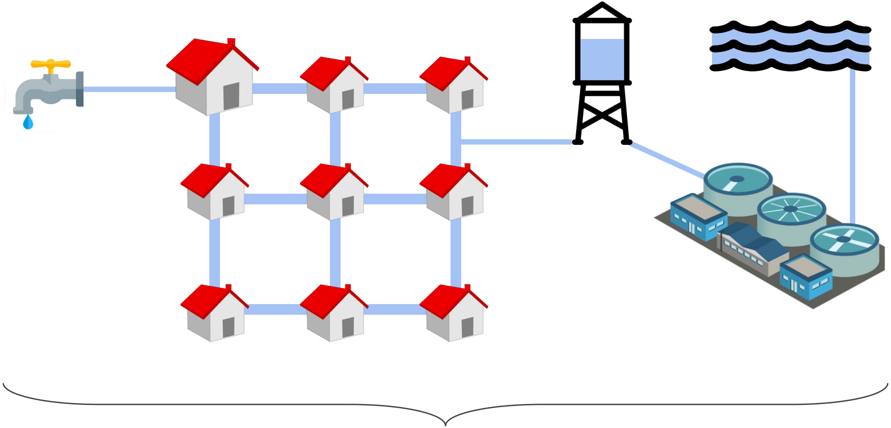
- Challenge: If we only use location as independent variables, the Bayes decision boundary is very complex!
- But if we use other independent variables (a simple one would be local emissions) we might do ok.

Clasification application example: Predicting water quality violations

Seigi Karasaki, last year's ER131 GSI, has a work in progress to predict water quality violations in small community water systems in California.

He kindly shared the next several slides with me.

Our relationship with water



Community water system

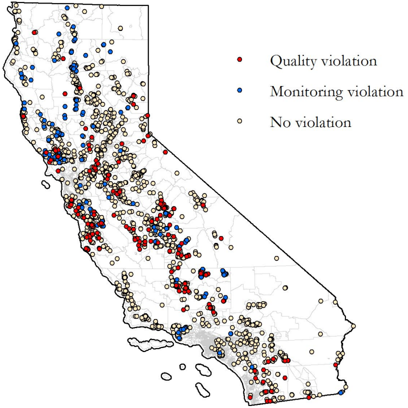
Environmental justice and drinking water

<u>2012:</u>

California recognizes a human right to "safe, clean, affordable, and accessible" drinking water (AB685).

<u>2017:</u>

∼300 systems across the state lack access to safe drinking water. Disadvantaged communities are disproportionately affected and at risk.



The burden of providing clean water rests with community water systems, but not all systems have the capacity to monitor for, let alone provide, clean and safe drinking water.

These systems are often “hidden” from the state, and are *de facto* excluded from assistance programs.

Research Question

Can we predict water quality violations for small community water systems that have shown lapses of technical and managerial capacity in the last five years?

Approach

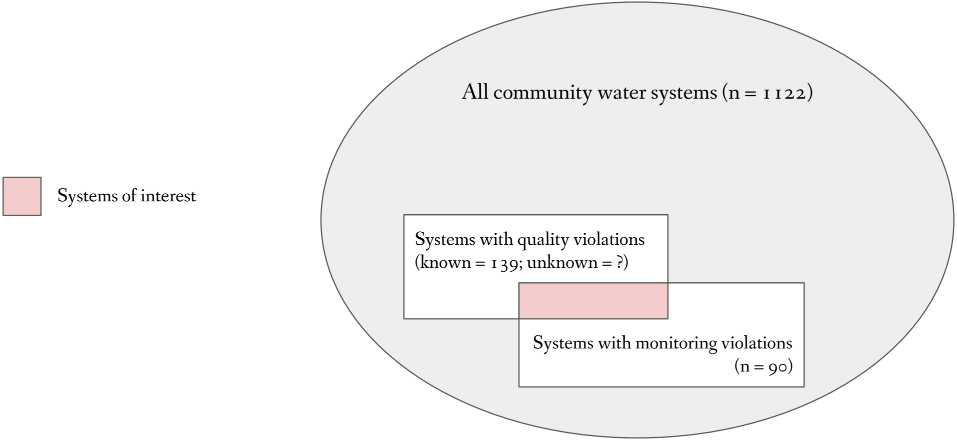
1. Describe the relationship between water quality/monitoring violations and system-level demographic data;

2. Predict the likelihood of water quality violations within identified capacity-deficient systems.

Data preparation

1. Basic water system information small community water systems in California ($n = 1122$)
    - Geospatial data, water source type, connection count, population count, county
    - Small = piped connections < 200 & population served < 10,000
2. Approximated demographic information for each community water system
    - American Community Survey (2012); processed via ArcMap, R
3. Water quality violation data ($n = 139$)
    - Human Right to Water Portal (State Water Resources Control Board)
4. Water monitoring violation data ($n = 90$)
    - Manually pulled from 2646 pdf letters (Enforcement actions, State Water Resources Control Board)

Identifying "hidden" systems



Spatial distribution of “hidden” violations

Predicting water quality violations
for capacity-deficient systems

PWSID	NAME	SCORE	COUNTY
CA1000019	EL PORVENIR	1	FRESNO
CA1000021	WILDWOOD ISLAND	1	FRESNO
CA1000247	SHERWOOD FOREST MHP	1	FRESNO
CA1000359	CANTUA CREEK	2	FRESNO
CA1000546	FIVE POINTS	2	FRESNO
CA1300009	WINTERHAVEN COUNTY WATER DISTRICT	1	IMPERIAL
CA1500405	AERIAL ACRES WATER COMPANY, INC.	2	KERN
CA1500571	LUCKY 18 ON ROSAMOND, LLC	1	KERN
CA1500578	LONG CANYON WATER COMPANY CORP.	2	KERN
CA3301528	OAK HAVEN ASSOCIATION	1	RIVERSIDE
CA3500006	STONEGATE WS	1	SAN BENITO
CA4300630	FOOTHILL MUTUAL WATER	1	SANTA CLARA
CA4900603	RIEBLI MUTUAL WATER COMPANY	1	SONOMA
CA4900901	SOBRE VISTA WATER COMPANY	1	SONOMA
CA5400710	BADGER HILL ESTATES	1	TULARE