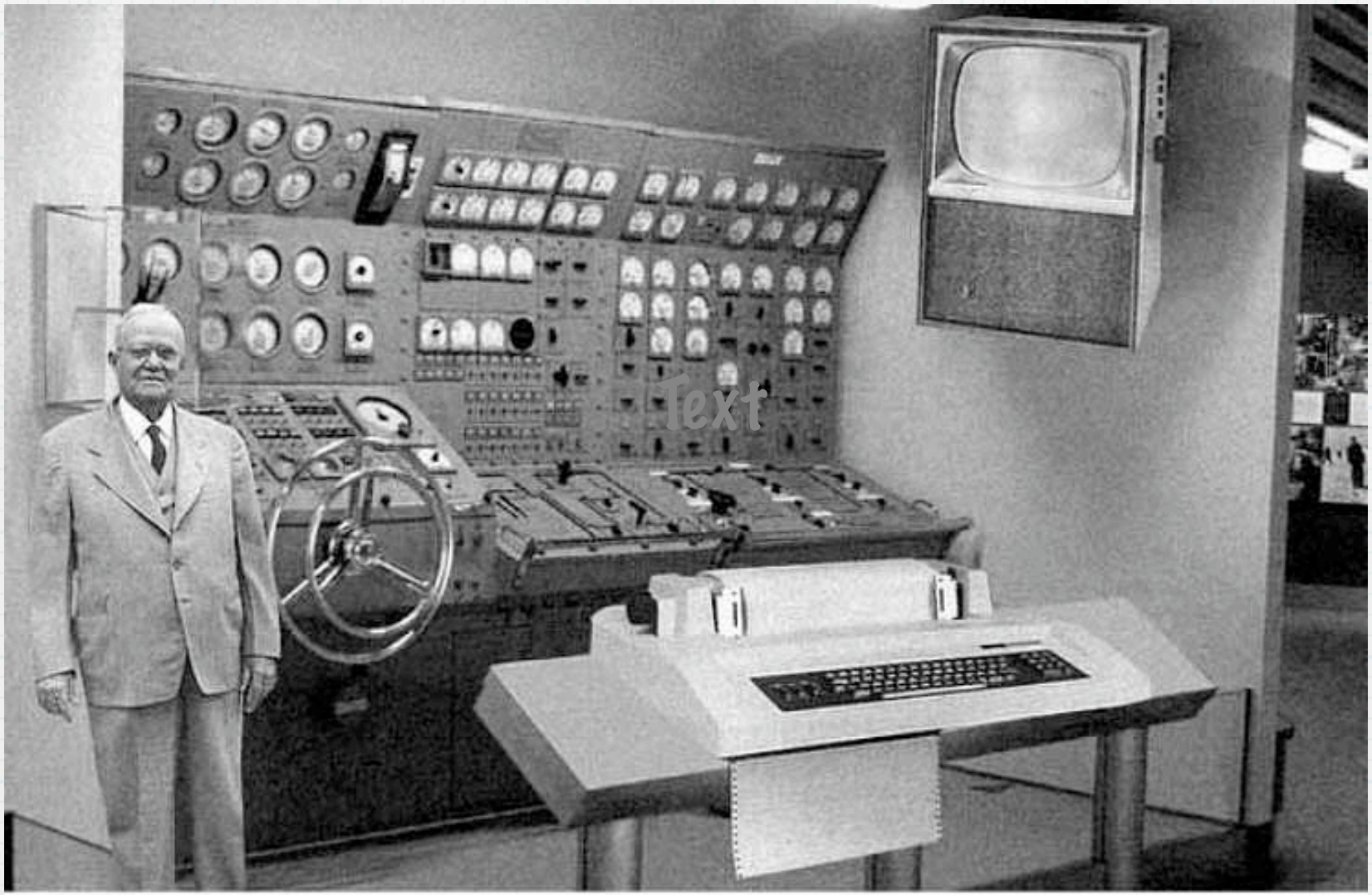
Forecasting Murder and Other Serious Crimes

Richard Berk
Department of Statistics
Department of Criminology

University of Pennsylvania

ANOTHER FORECAST - CAVEAT EMPTOR



Scientists from the RAND Corporation have created this model to illustrate how a "home computer" could look like in the year 2004. However the needed technology will not be economically feasible for the average home. Also the scientists readily admit that the computer will require not yet invented technology to actually work, but 50 years from now scientific progress is expected to solve these problems. With teletype interface and the Fortran language, the computer will be easy to use.

To begin, we rounded up all of the usual methodological suspects...

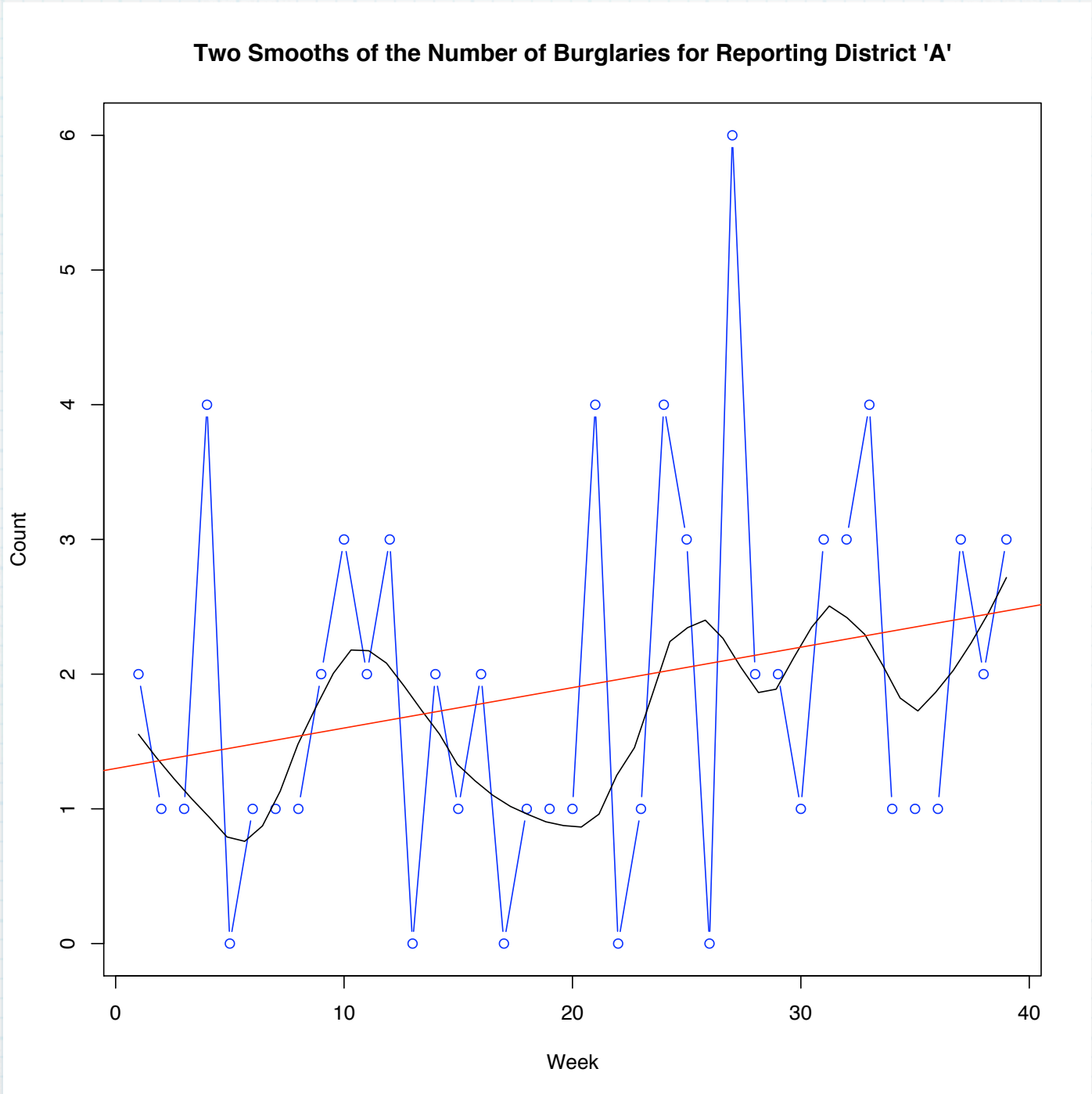
* Using LAPD offense data, define spatial and temporal units.
* Locate crimes in time and space.
* The response (Y) is the number of crimes (of different types) with predictors (X) lagged values of various sorts and some other predictors such as weather.
* Fit a stochastic model of the form Y=f(X) + Noise.
* Extrapolate that model into the future along with its uncertainty.

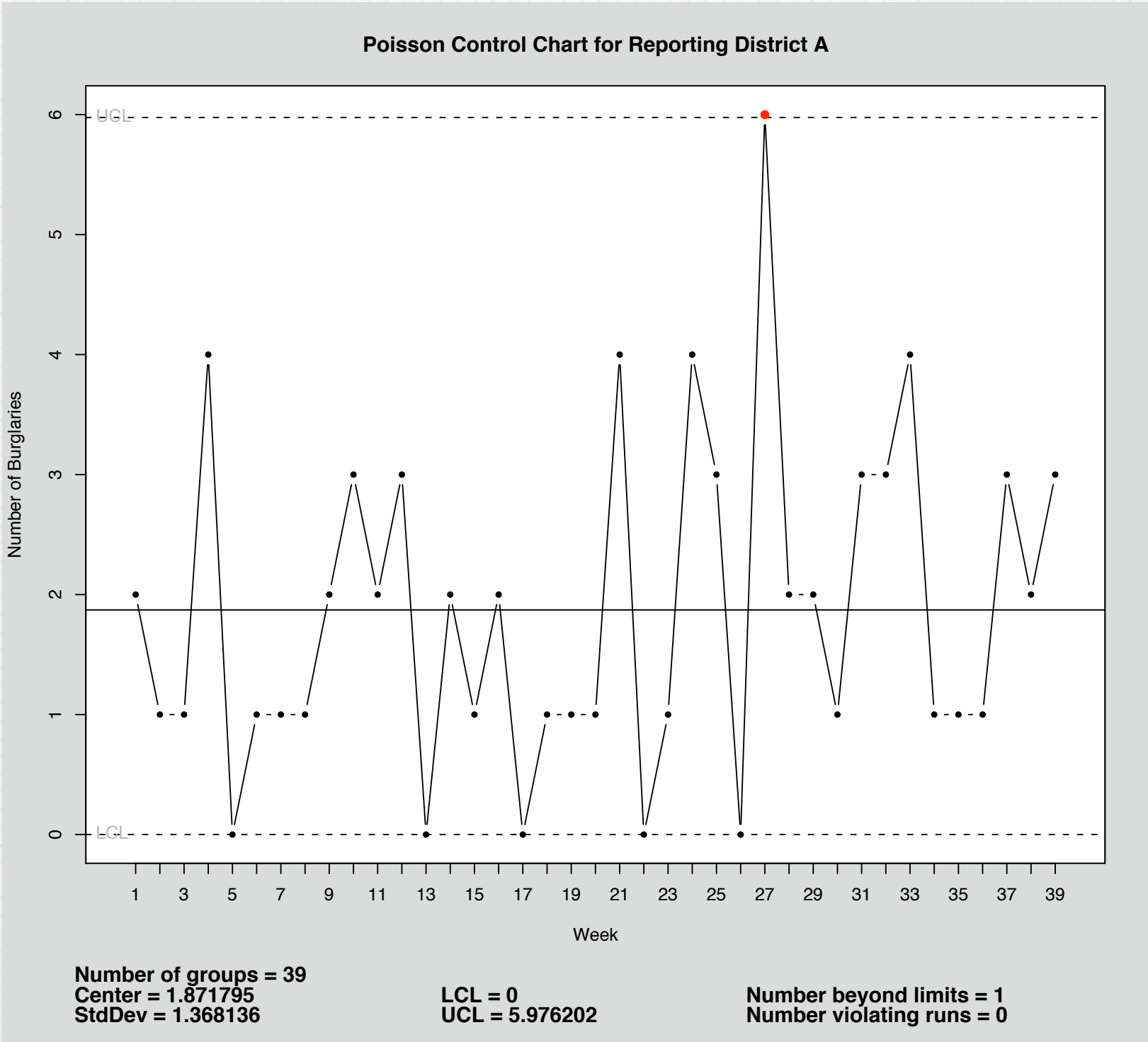
Some Practical Difficulties

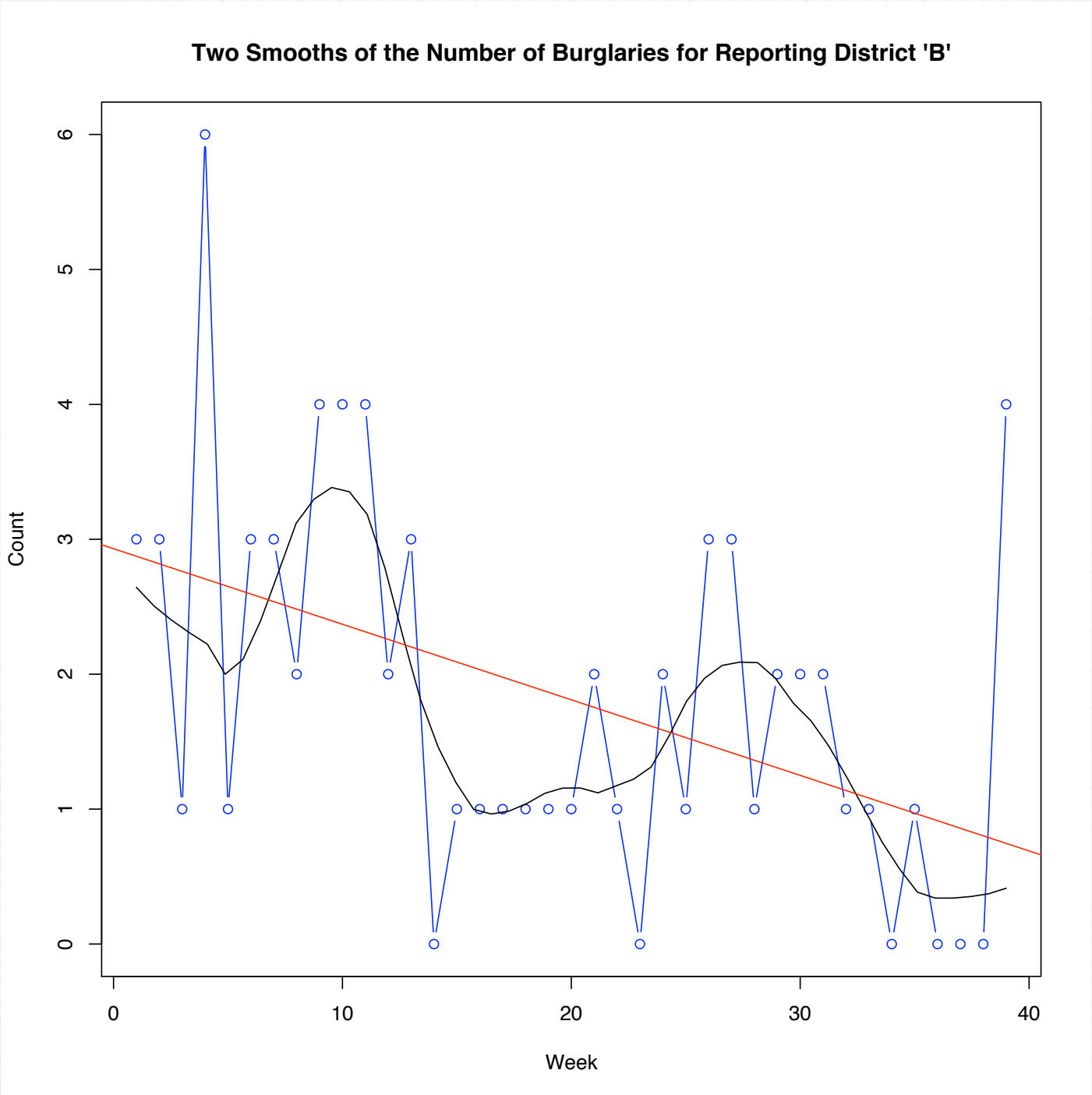
* Spatial units smaller than reporting districts are probably preferable for operational reasons (at least), but then the crime events are relatively rare.
* Serious crimes are more rare still.
* Crime categories often don't map that well on to behavior.
* Many crimes go unreported.
* A very weak signal can result from little variability and a large chance component (e.g., homicides).

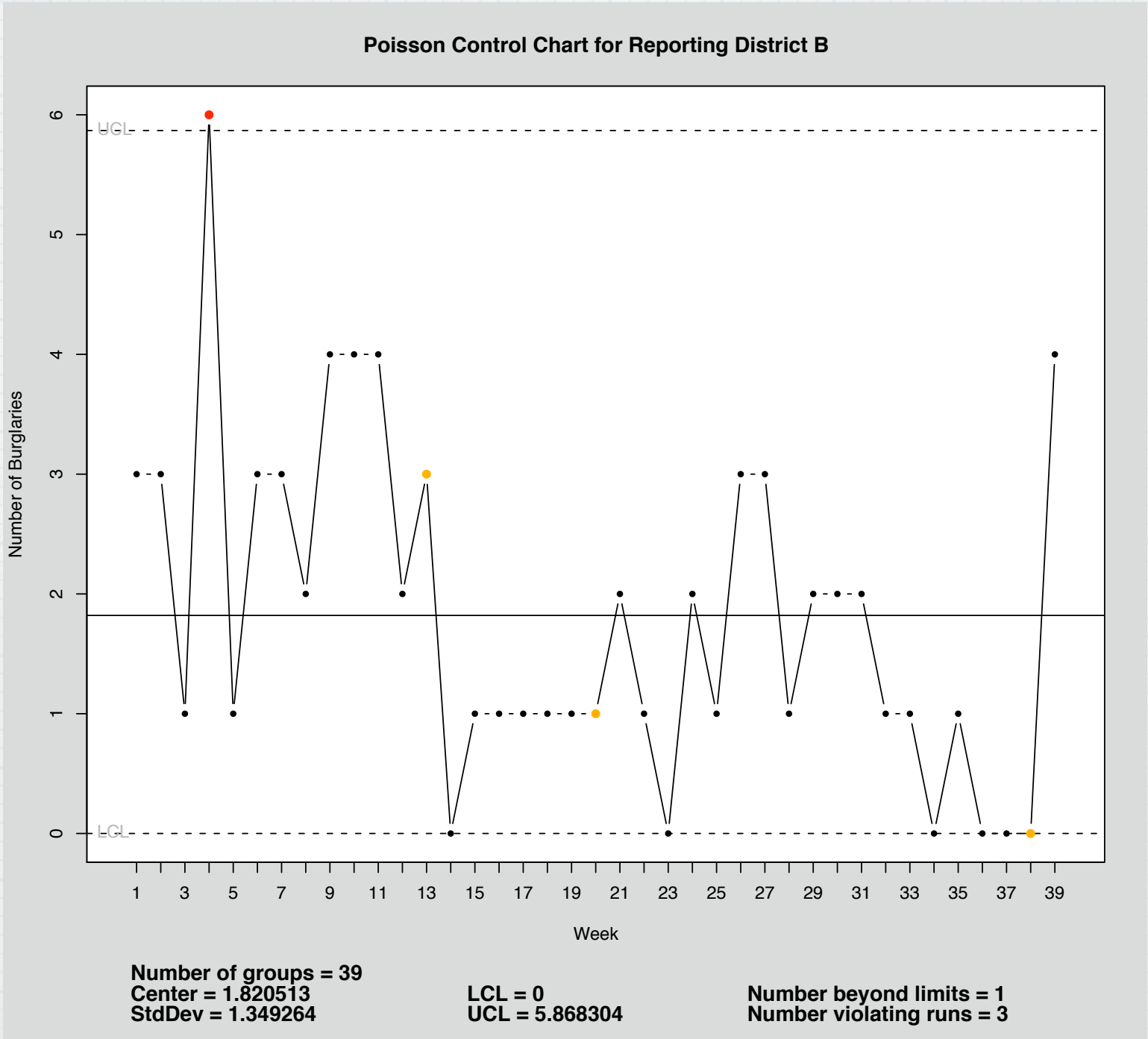
So, let's look at some illustrations...

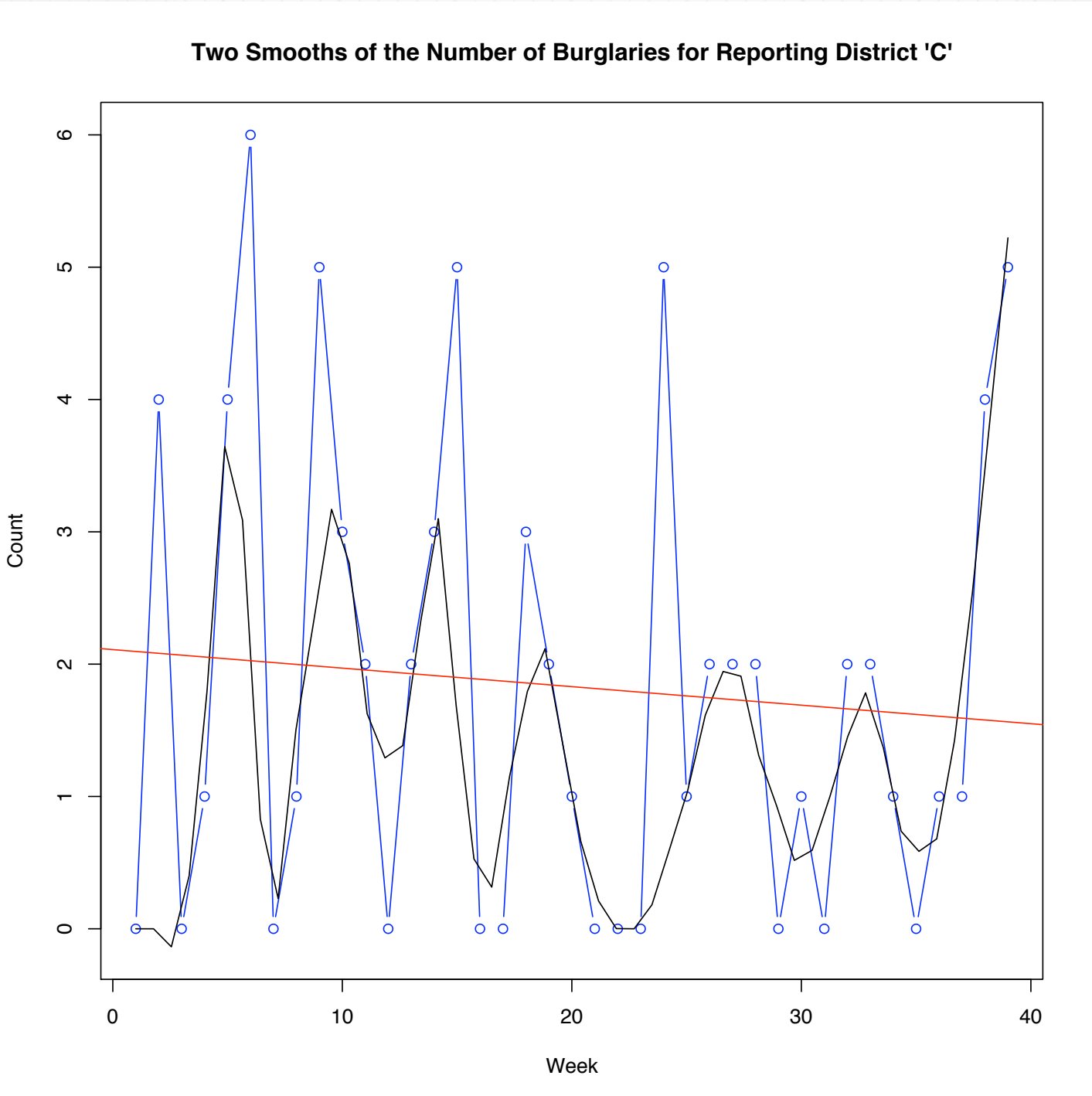
* Burglaries.
* Three separate but adjacent LAPD Reporting Districts.
* Chosen because burglaries are relatively common and might show a temporal signal.
* 39 weeks with counts aggregated up to the week (based on COMPSTAT experience).
* Can we see something that looks like temporal hot spots --- concentrations of crime beyond the usual range of variation?

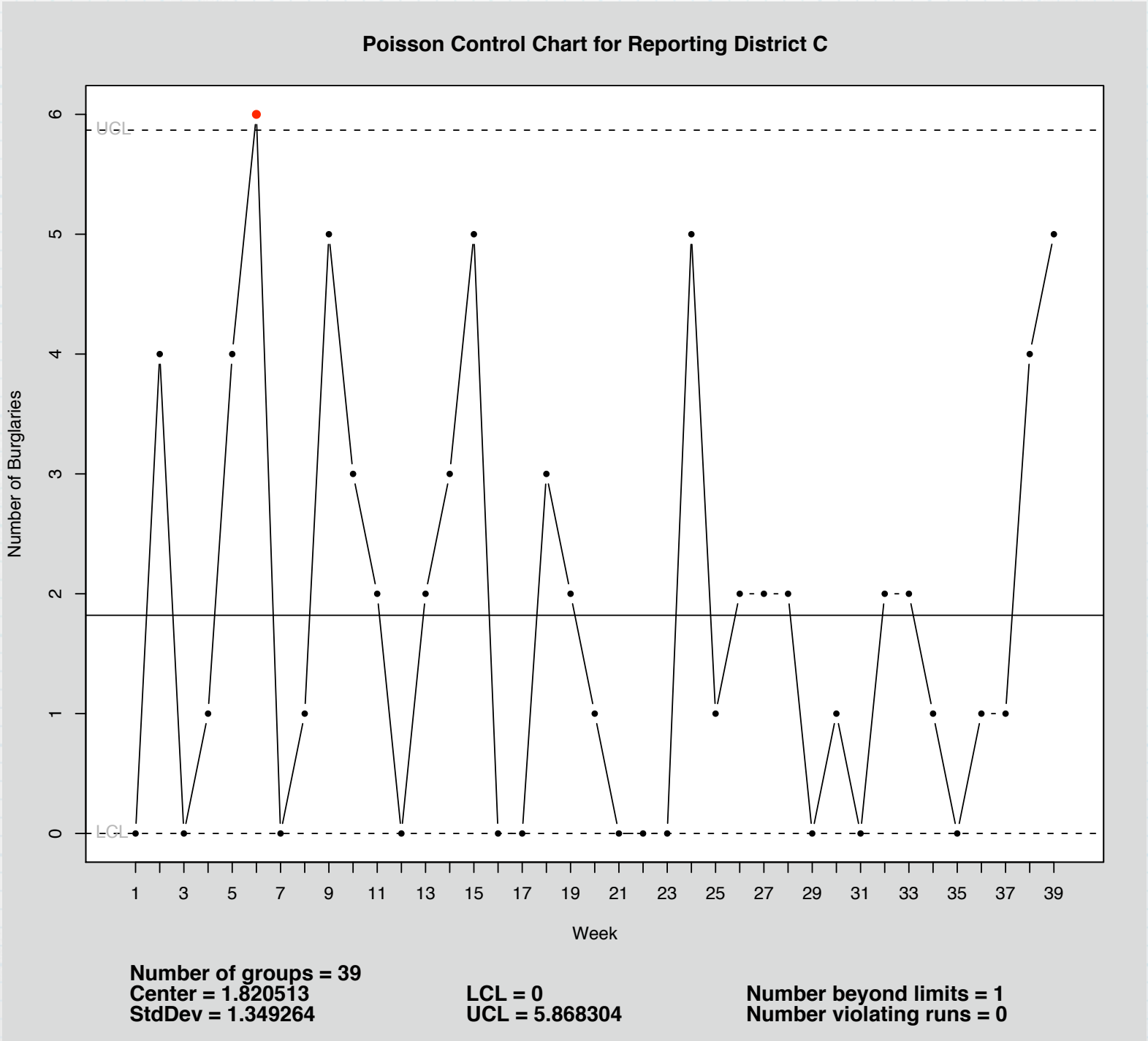

Two Smooths of the Number of Burglaries for Reporting District 'A'
Count
0
1
2
3
4
5
6
0
10
20
30
40
Week


Poisson Control Chart for Reporting District A
UCL
LCL
Number of Burglaries
0
1
2
3
4
5
6
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
Week
Number of groups = 39
Center = 1.871795
StdDev = 1.368136
LCL = 0
UCL = 5.976202
Number beyond limits = 1
Number violating runs = 0


Two Smooths of the Number of Burglaries for Reporting District 'B'
Count
Week
0
1
2
3
4
5
6
0
10
20
30
40


Poisson Control Chart for Reporting District B
Number of Burglaries
Week
UCL
LCL
0
1
2
3
4
5
6
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
Number of groups = 39
Center = 1.820513
StdDev = 1.349264
LCL = 0
UCL = 5.868304
Number beyond limits = 1
Number violating runs = 3


Two Smooths of the Number of Burglaries for Reporting District 'C'
Count
0
1
2
3
4
5
6
0
10
20
30
40
Week


Poisson Control Chart for Reporting District C
Number of Burglaries
Week
UCL
LCL
0
1
2
3
4
5
6
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
Number of groups = 39
Center = 1.820513
StdDev = 1.349264
LCL = 0
UCL = 5.868304
Number beyond limits = 1
Number violating runs = 0

Some Conclusions

* Mean is about the same as the variance in each reporting district, and a Poisson formulation with a common mean across all three reporting districts looks pretty good.
* Not much forecasting help there.
* Risk is that personnel allocation decisions will be made on chance patterns followed by regression to the mean.
* Large spatial variation that is pretty stable: South Central v. Sherman Oaks --- No news there and no help to the LAPD.
* Some temporal patterns for the city as a whole that look like the sorts of time series models you'd expect --- also no news and no help to the LAPD.
* We are still working on this and have some ideas that might help, but we also need to think about a different approach.

Another way....

* The goal is to map crimes that have not yet been committed in time and space.
* Forecast which individuals are high risk for various kinds of crimes.
* Locate them in time and space (because it is common for people to commit crimes in their own neighborhoods).
* Aggregate risk to those temporal and spatial units.

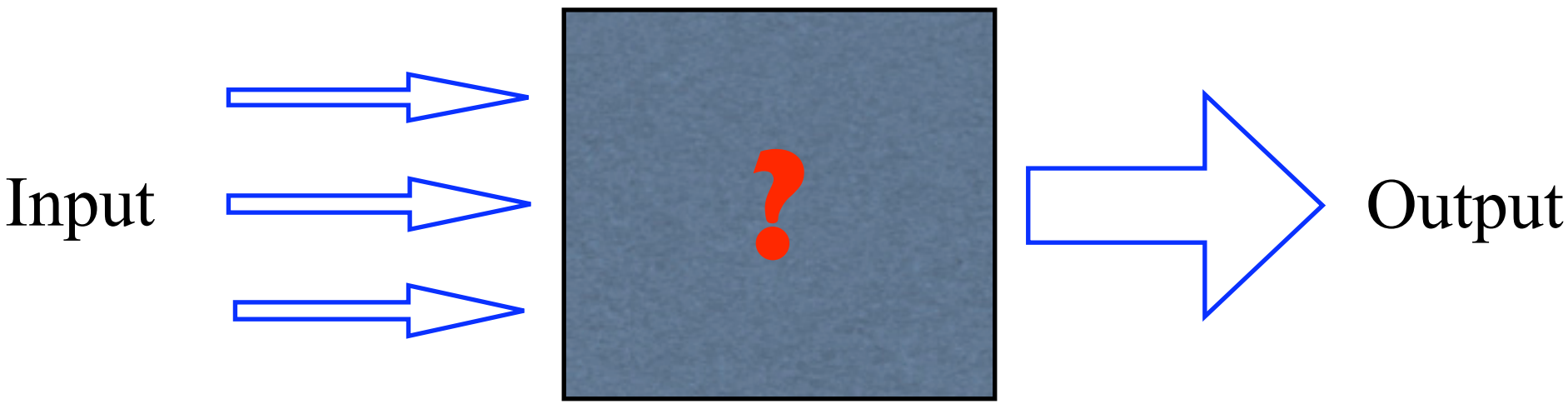
The Philadelphia Story

* High and increasing number of homicides in Philadelphia --- up 20% past two years.
* Perhaps a third committed by individuals on probation or parole.
* Could individuals likely to commit a homicide be given more intensive services while on probation in order to reduce risk?
* So,... we need a practical forecasting tool as a case comes in the front door.
* With this in hand, we can also locate high risk cases by addresses that can in turn affect police resource decision --- we map “future homicides.”
* And we can pass the names along to detectives as possible suspect when neighborhood homicides occur.

Some Background on the Approach

* Outcome is homicide or attempted homicide within 2 years of being assigned a probation officer.
* About 1% fail by this criterion (eliminates most conventional statistical procedures).
* Predictors drawn from routine administrative data.
* Using statistical learning procedures.
* True forecasting skill as the gold standard.
* Take the costs of false positives and false negatives into account.

No-Apology Black-Box Modeling



Forecasting accuracy is the gold standard.

Random Forests

1. Take a random sample of size N with replacement from the data. (This is a training sample. Those not selected are a test sample.
2. Take a random sample without replacement of the predictors.
3. Construct the first CART partition of the data.
4. Repeat step 2 for each subsequent split until the tree is as large as desired. Do not prune.
5. Drop the test sample down the tree. Store the class assigned to each test sample observation along with each observation's predictor values.
6. Repeat steps 1-5 a large number of times (e.g., 500).

Continue

7. Using only the class assigned to each observation when that observation is not used to build the tree, count the number of times over trees that the observation is classified in one category and the number of times over trees it is classified in the other category.

8. Assign each case to a category by a majority vote over the set of trees.

Assets

* Helps significantly to control overfitting.
* Effectively no limit to the number of predictors
* Provides an effective response to the curse of dimensionality.
* Will inductively determine non-linear relationships.
* Will allow highly specialized predictors to participate.
* Will find highly non-linear relationships, even step functions.
* Will allow for asymmetric costs of false positives and false negatives.

The Data

* For Philadelphia
* For the entering cohort of 2002.
* For a two-year follow-up.
* 66,518 cases.
* 100's of potential predictors.
* 737 cases charged with a homicide or attempted homicide (1.1%).
* We use a random sample of 30,000 cases and 30 predictors.
* Roughly a 1 to 10 cost ratio (FP/FN)

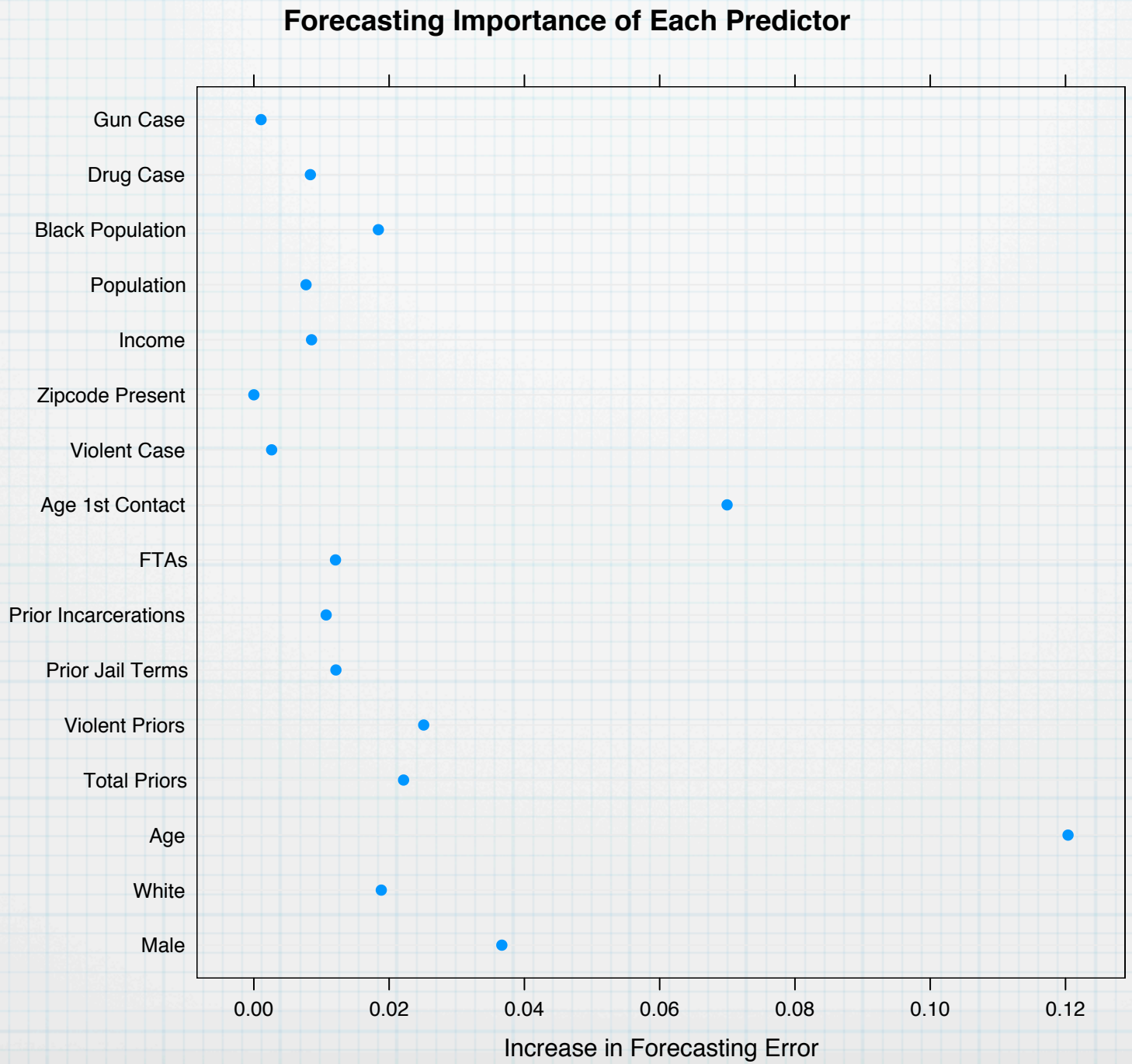
Forecasting Skill

* 93% accurate overall, But...
* 94% accurate for true negatives.
* 46% accurate for true positives.
* About 10 false positives for 1 true positive --- without the model it would be 100 to 1.
* In practice, about 10% are classified as high risk

Let’s talk about false positives.

* They are almost inevitable in practice.
* False positives can be traded against false negatives.
* Kinder and gentler services can be offered.
* Low-risk model soon to “net-shrink.”
* Nevertheless, a political problem.

Forecasting Importance



The key is less the predictors that surface (no surprise there) and more how the information in the predictors is used.

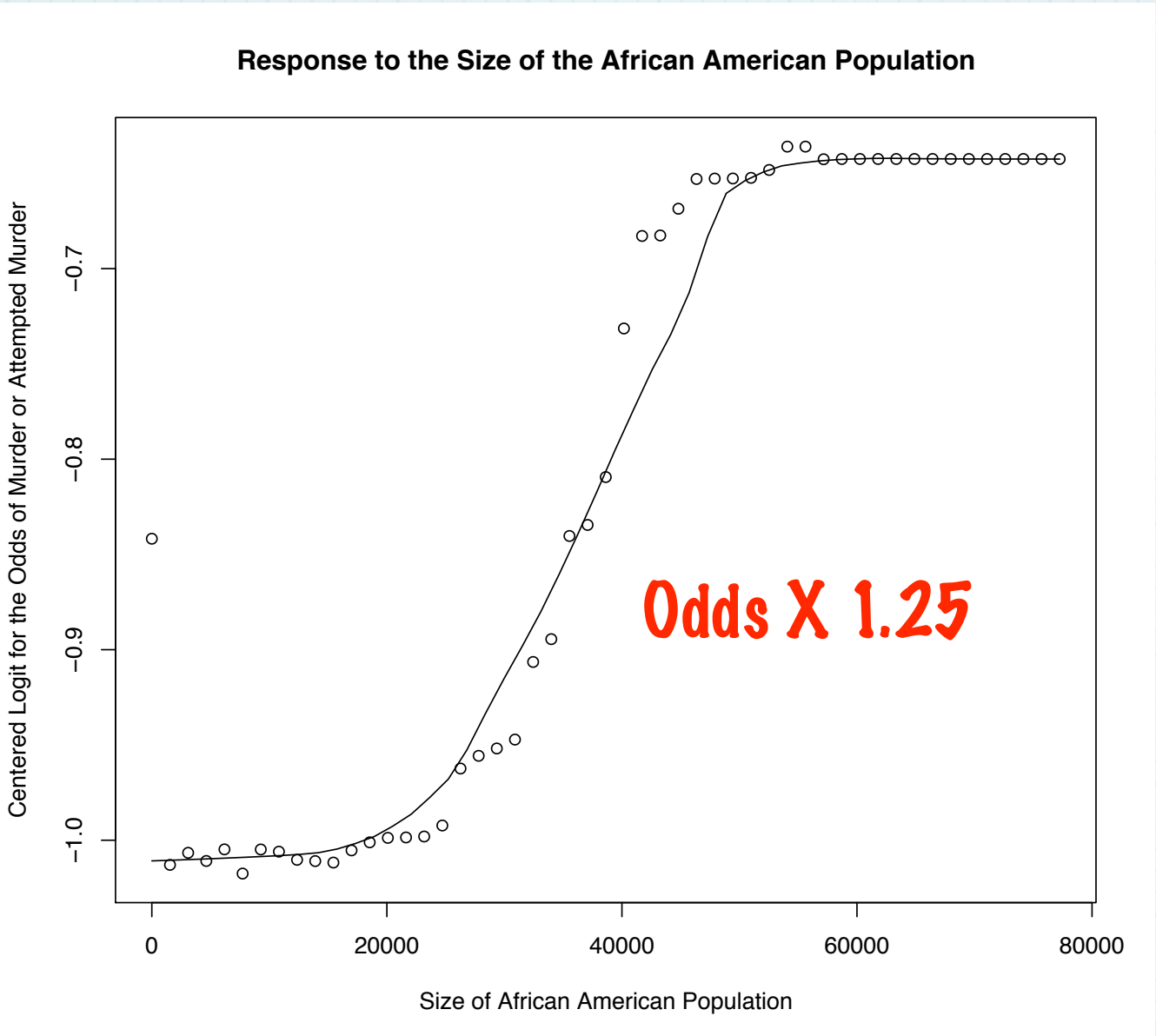
Response to Age

Response to Age at First Contact

Response to the Number of Violent Priors

Response to the Total Number of Priors

Response to the Size of the African American Population



At the end of the day...

* For the total pool of probationers, 1 in 100 commit a homicide or attempted homicide.
* For our high risk pool of probationers, 1 in 10 commit a homicide or attempted homicide.
* They key statistical point is that we can capitalize inductively on highly non-linear relationships without overfitting.

What Next?

* We have refined the model beyond what I have shown.
* We can update quickly (e.g., once a week) with new cases as things change.
* We have our first cohort of high risk cases.
* We have started an RCT for a random sample.
* We are mapping areas at high risk for homicide beyond the obvious.
* We are moving toward cooperating with the DA and homicide detectives.
* We are seeking data and cooperation for juvenile authorities.
* We are well on our way toward a low risk model so that resources can be better allocated.

End

The Log-Odds Metric

$$f_k(X) = \log[\mathrm{p}_k(X)] - \frac{1}{K}\sum_{k=1}^{K}\log[\mathrm{p}_k(X)].$$