Geo-spatial Analysis with Generalized Additive Models

CAS Annual Meeting
Chicago
November, 2011

Jim Guszcza
Deloitte Consulting LLP
The University of Wisconsin-Madison

Antitrust Notice

The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.

Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.

It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

Agenda

Spline Regression Recap

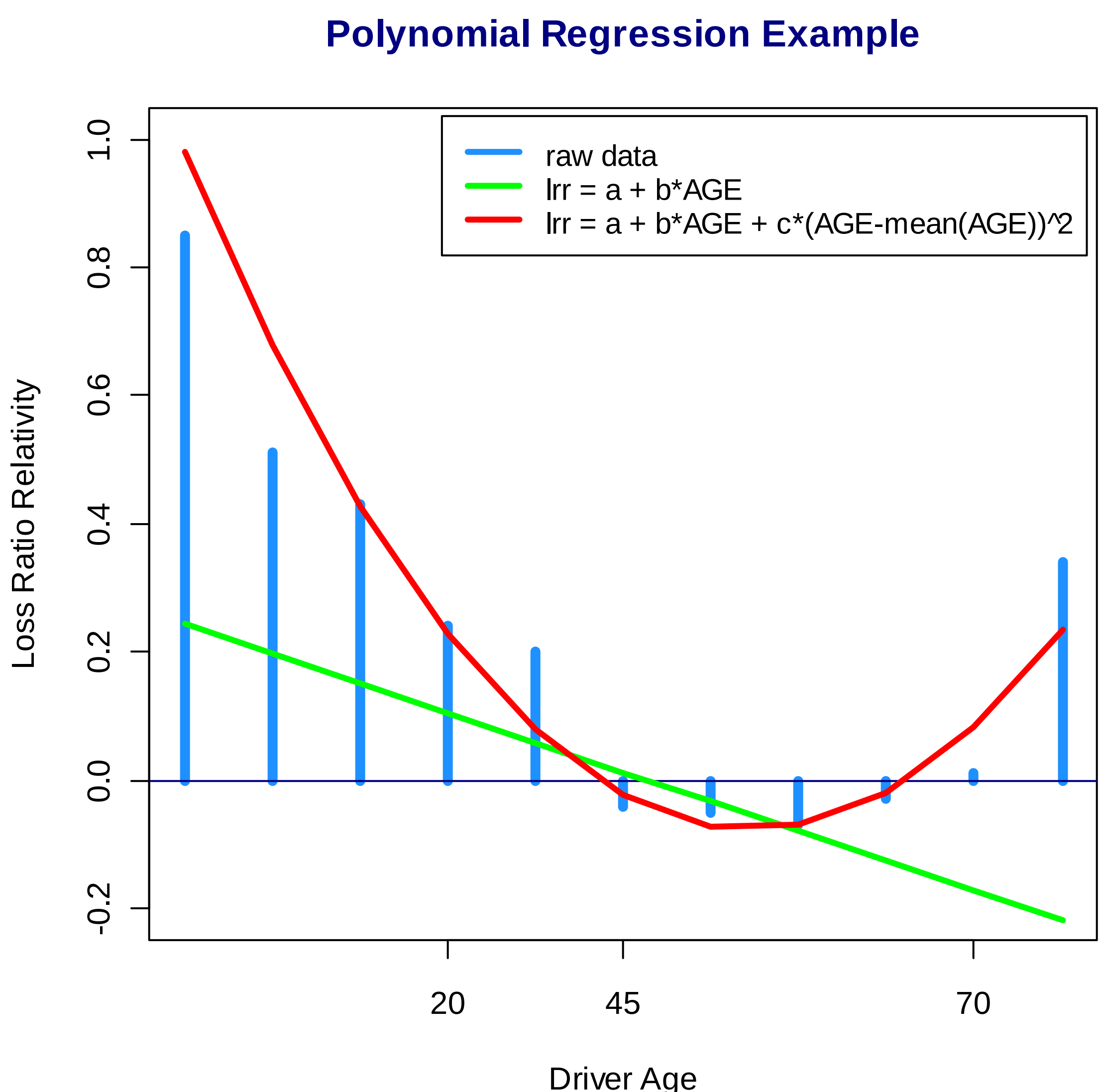
Generalized Additive Modeling Theory

Geo-spatial GAM example

Spline Regression

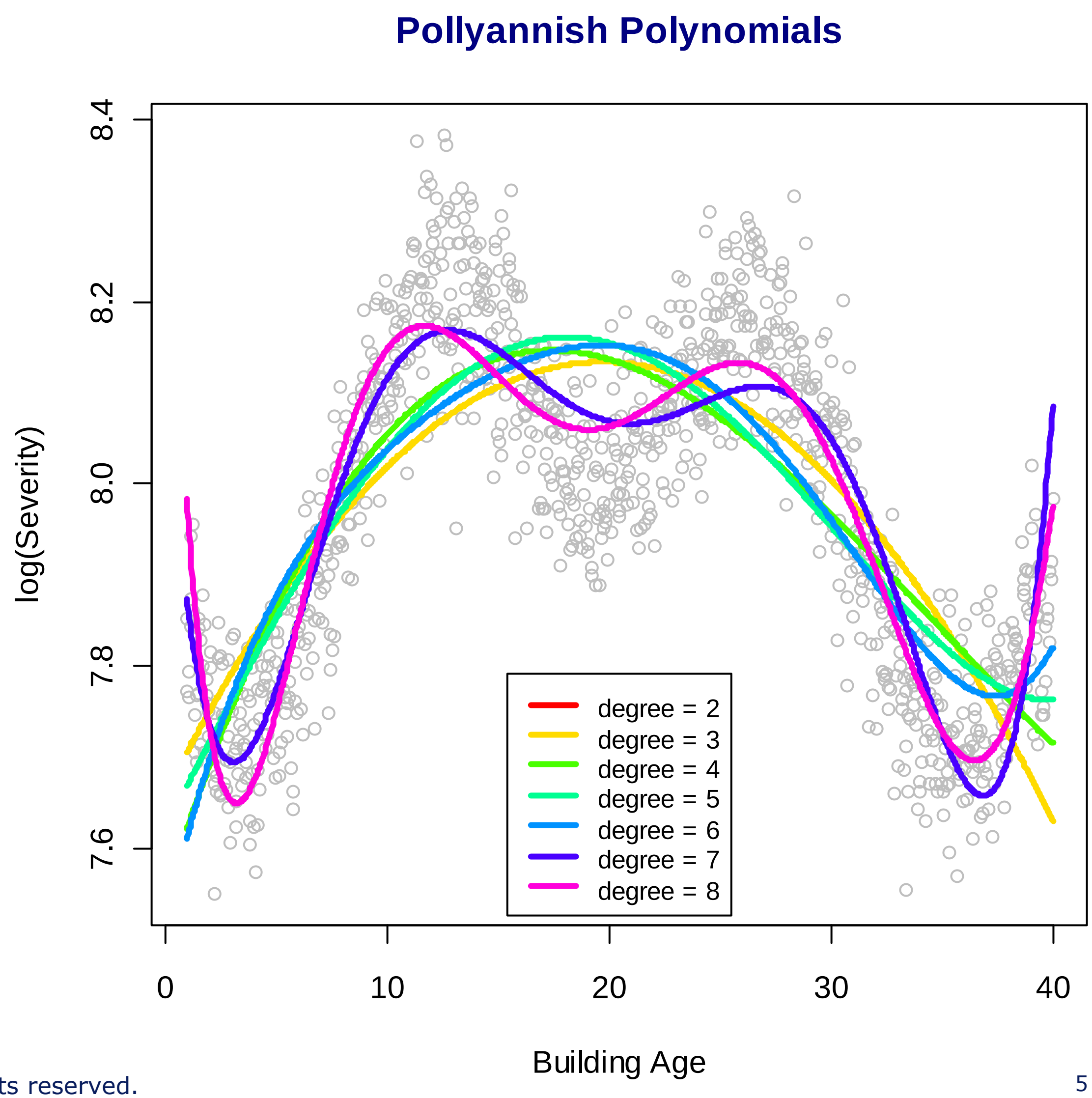
Modeling Non-Linear Patterns

- Linear models only have to be linear in the parameters.
- By cleverly transforming our variables we can model just about any non-linear relationship.
- Often in practice, adding a quadratic and maybe cubic terms will suffice.
- Here, adding a quadratic term results in a reasonable fit.



The Limits of Polynomial Regression

- In more complex cases, adding polynomial terms is not enough.
- This (exaggerated) example illustrates the limitations of polynomial regression.
- Adding quadratic and cubic terms is better than nothing, but doesn't fully capture the pattern.
- Even an 8^{th} degree polynomial regression provides only a rough approximation.

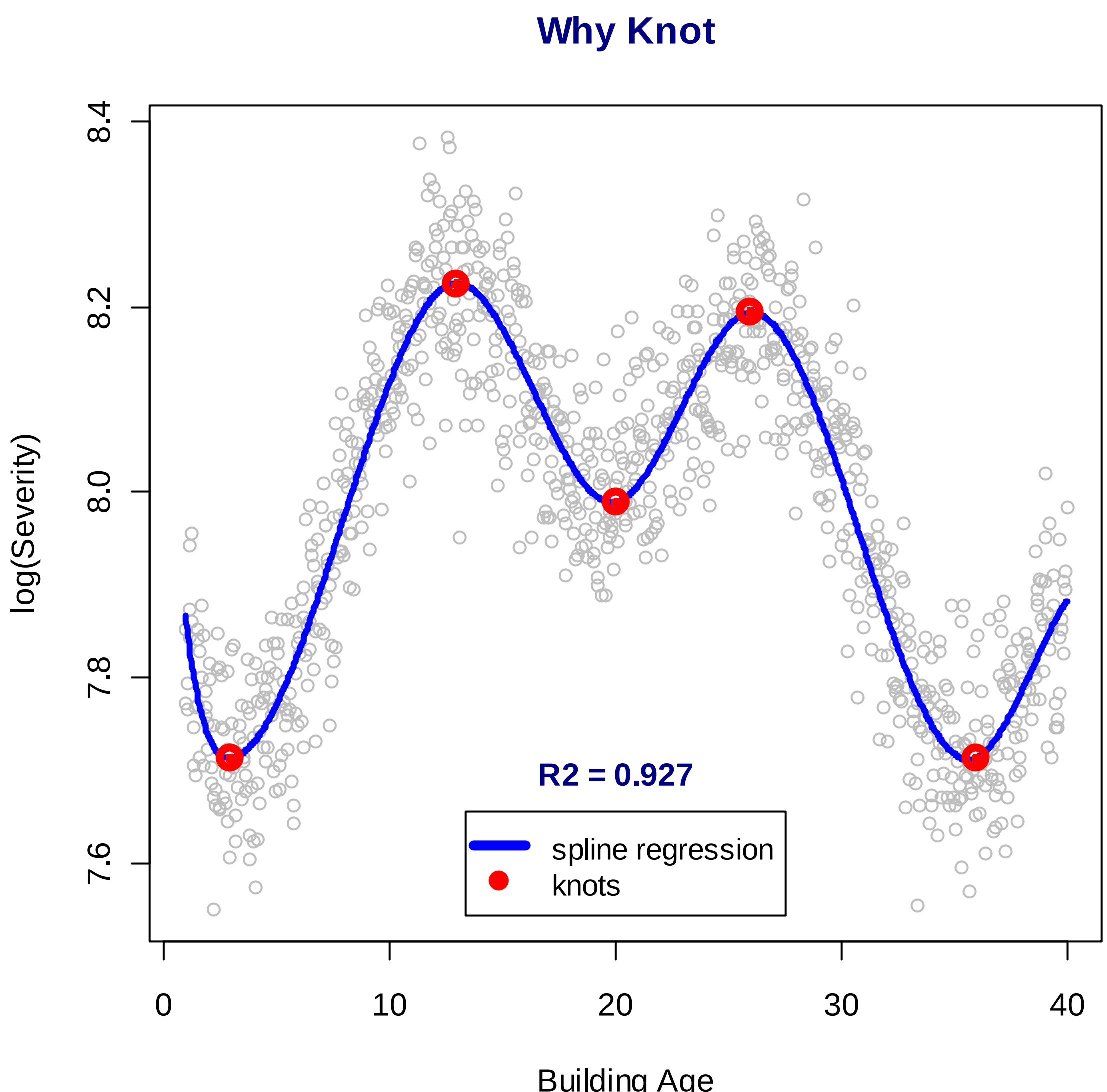


Cubic Spline Regression

- In more complex cases such as this, cubic spline regression is an excellent alternative.

- Here we have a series of cubic polynomials joined at a series of manually selected knots.
    - The model is “smooth” in the sense that it has continuous 1st and 2nd derivatives at each knot.

- In this case, a cubic spline regression with 5 knots achieves an excellent fit (R^2=0.93).



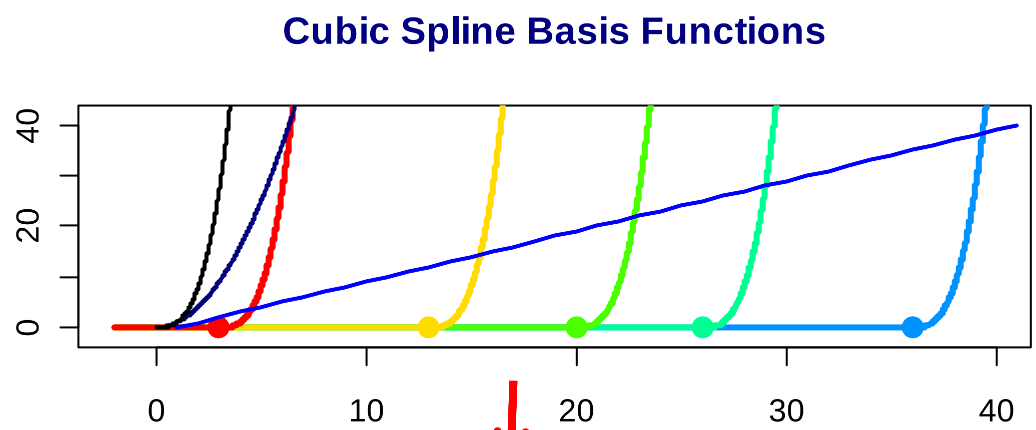
Basis Basics

- The basic trick is to identify a collection of **basis functions** $\{b_i(x)\}$ that can approximate any functional form.

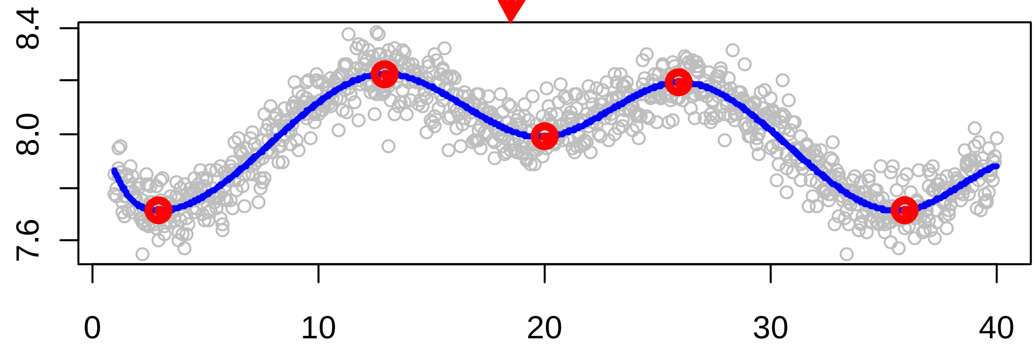
- In addition to polynomial terms, our spline regression includes a linear combination of these basis functions of building age:

$$b_{k[i]}(x) = \begin{cases} (x-k)^3 & x > k \\ 0 & x <= k \end{cases}$$

  - Aside: the "hockey stick functions" used in the MARS algorithm are the lower-degree analog of these basis functions.

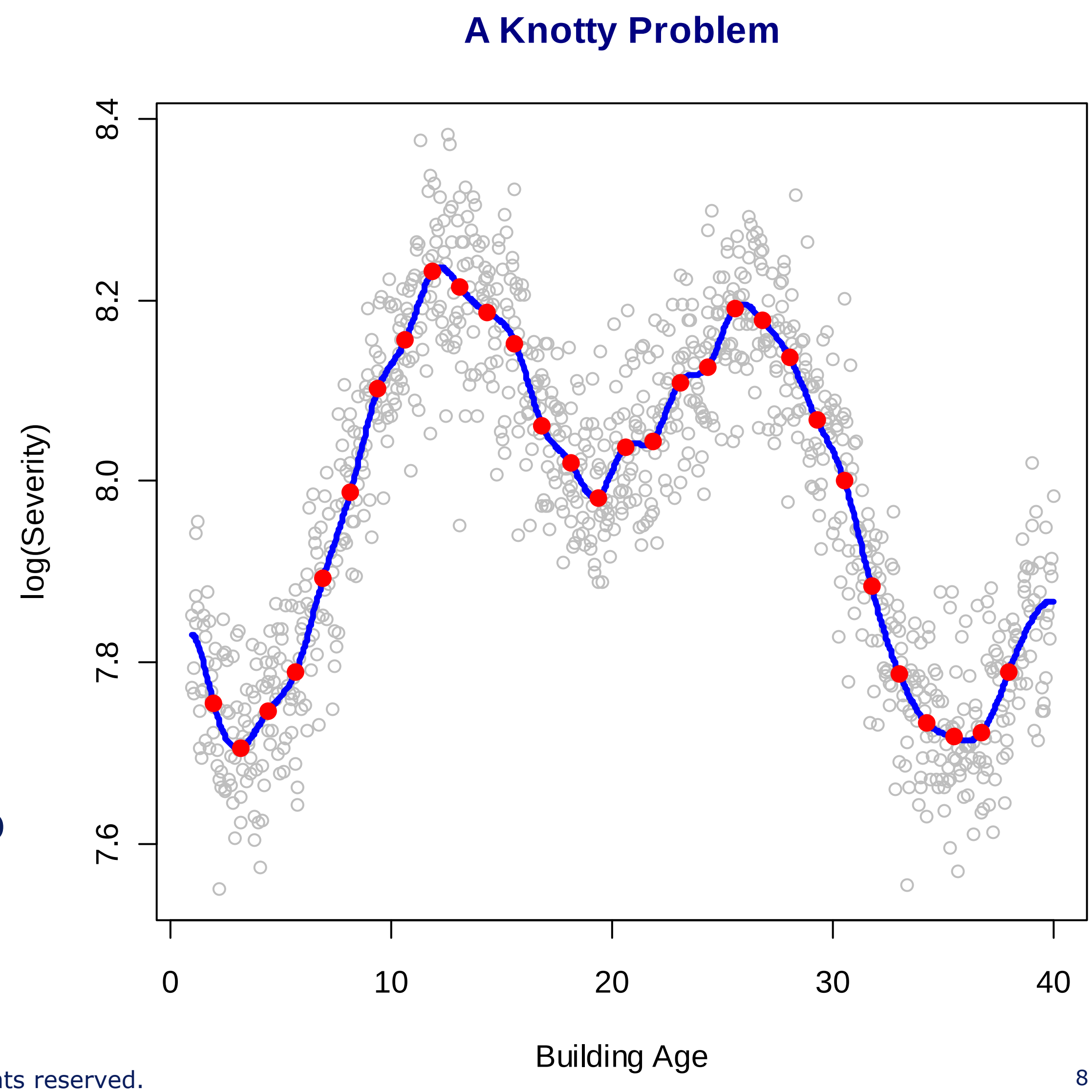


$$f(x) = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{i=1}^{k} \gamma_i * b_i(x)$$



Overly Caffeinated Spline Regression

- Spline regression is great, but we must be careful when selecting the knots.
- Too few knots ➔ not all of the patterns will be reflected in the model.
- Too many knots ➔ our model will fit random noise in the data.
- Capturing too much random noise can lead to a model that performs poorly out-of-sample.
  - We'll come back to this point.



Generalized Additive Models

Generalized Additive Models

- Recall the basic ideas of Generalized Linear Models:
  1. $g(\mu) \equiv g(E[Y]) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_N X_N$
  2. $Y|\{X\} \sim$ exponential family

- Generalized <u>Linear</u> Models: $g(\mu)$ = linear combination of predictors

- Generalized <u>Additive</u> Models: the linear predictor can also contain one or more *smooth functions* of covariates.

$$g(\mu) = \beta \cdot \boldsymbol{X} + f_1(X_1) + f_2(X_2) + f_3(X_3, X_4) + \ldots$$

- Note that some of the *f* can be functions of more than one predictor.

- This brings us a lot of flexibility… but we need to figure out how to represent the functions $\{f\}$.

Generalized Additive Models

- GAM form:

$$g(\mu) = \beta \cdot \boldsymbol{X} + f_1(X_1) + f_2(X_2) + f_3(X_3, X_4) + \ldots$$

- How do we represent the functions $\{f\}$?
- Cubic splines offer an obvious answer.
- But recall that we had to choose the knot placements manually.
- This isn't good enough: we need a principled (and fairly automatic) way to specify a model that:
    - Fits the "true" linear and non-linear patterns in the data
    - But does not "over-fit" the data

Intuitively, it might seem that we need a way to determine the optimal placement of knots.

Fitting Signal, Not Noise

- **Alternate idea:** rather than worrying about which basis functions we need, we can fix the knots and basis functions ahead of time… but control the smoothness through <u>penalized least squares</u>.

- Rather than minimize SSE: $\sum_i \left(y_i - \sum_j \beta_j X_{ij} \right)^2$

- We can minimize ***penalized* SSE**: $\sum_i \left(y_i - \sum_j \beta_j X_{ij} \right)^2 + \lambda \cdot \int \left[f''(x) \right]^2 dx$

- The integral is a measure of the complexity of *f*(x).
  - Recall that our basis functions have continuous 2nd derivatives.

- The λ "smoothness" parameter determines how much we should penalize the complexity introduced by our cubic spline basis functions.
  - As $\lambda \rightarrow 0$, the GAM approaches an un-penalized regression spline
  - As $\lambda \rightarrow \infty$, the GAM approaches linearity

Penalized Least Squares

- The penalized SSE formula reflects a fundamental tradeoff.

$$\sum_i \left(y_i - \sum_j \beta_j X_{ij}\right)^2 + \lambda \cdot \int \left[f''(x)\right]^2 dx$$

More Basis Functions
Lower bias: Our spline model fits the data better ➔ 1st term is smaller.

Fewer Basis Functions
Higher bias: Our spline model fits the data worse ➔ 1st term is larger.

More Basis Functions
Higher Variance: there is a greater chance that the model will perform poorly out-of-sample ➔ 2nd term is larger.

Fewer Basis Functions
Lower Variance: there is a smaller chance that the model will perform poorly out-of-sample ➔ 2nd term is smaller.

- **This logic is sound… but we must determine the appropriate value of λ.**

Choosing λ

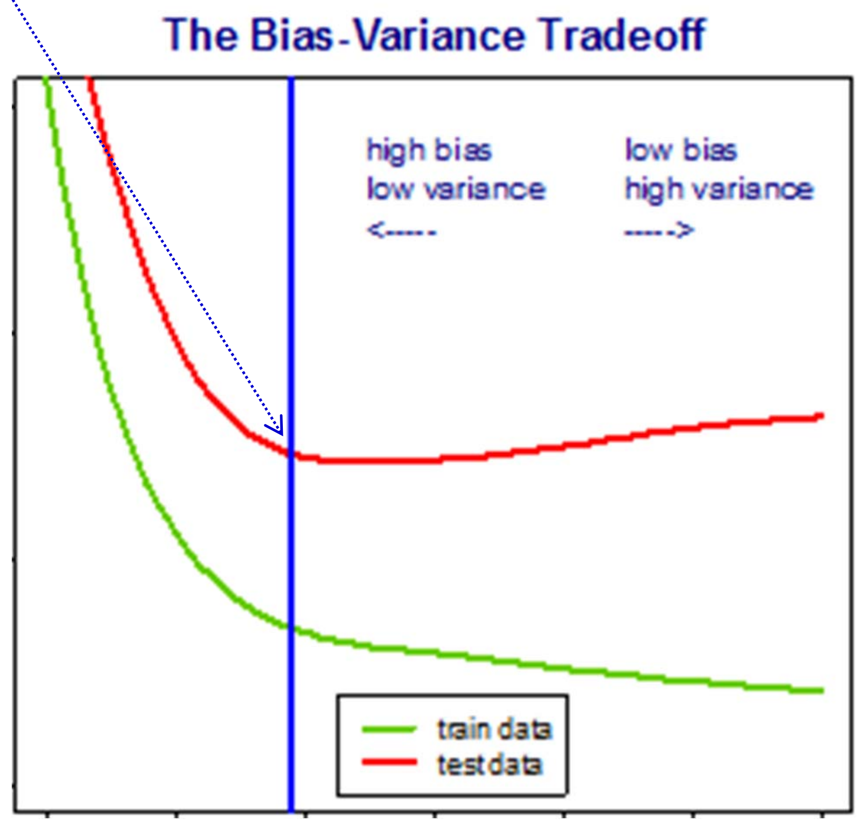
- We need a principles way to select λ before solving for the {β} parameters that minimize penalized SSE:

$$\sum_i \left(y_i - \sum_j \beta_j X_{ij}\right)^2 + \lambda \cdot \int \left[f''(x)\right]^2 dx$$

- We use **cross-validation** to do this.

- Select λ that minimizes SSE calculated using leave-one-out cross-validation.

- Conceptually the same idea used to determine the appropriate cost-complexity parameter in the CART algorithm.



To Summarize

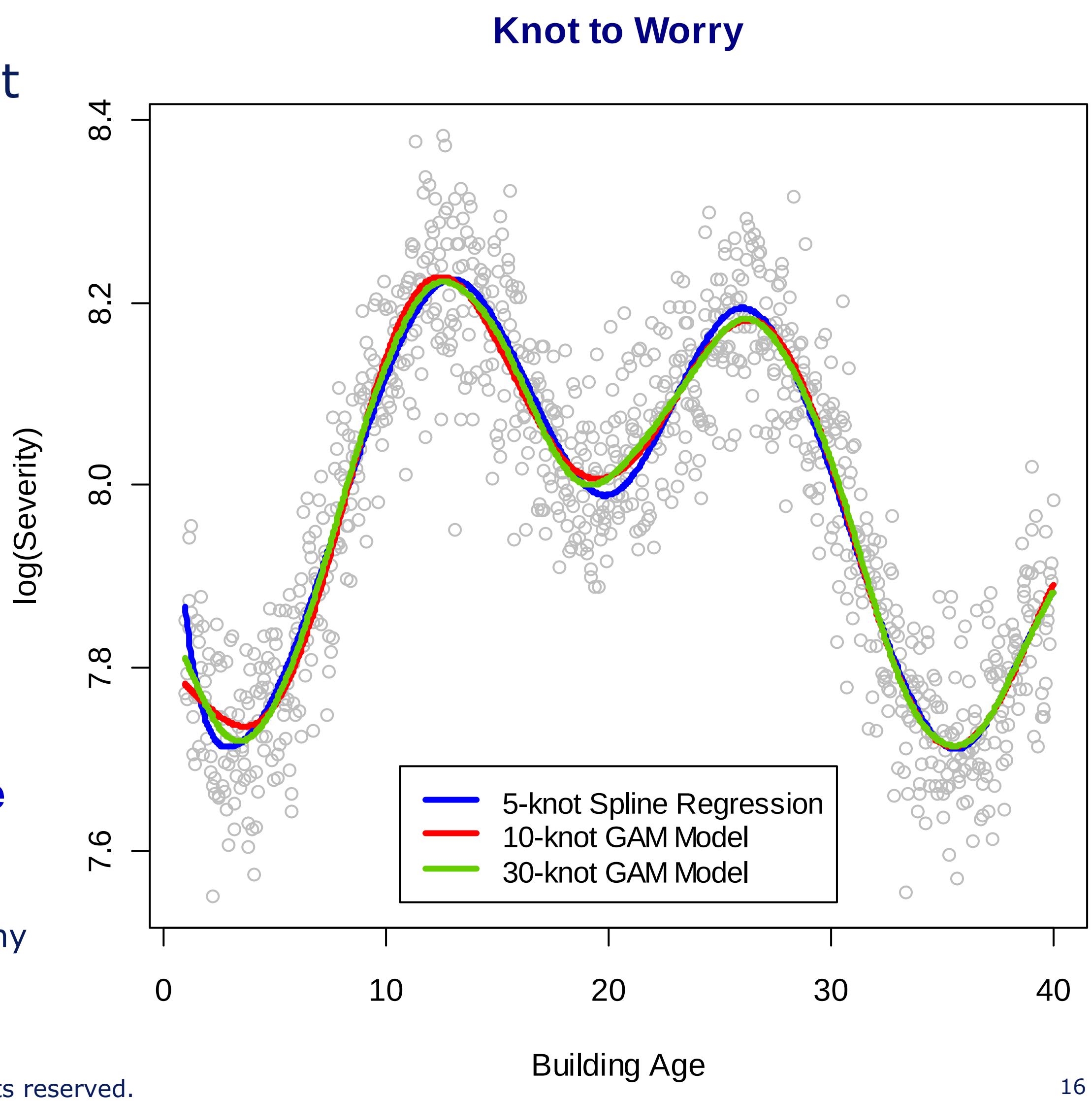
- Rather than manually select "just the right set" of knots and basis functions…
- We scatter the knots somewhat liberally…
- But add a 'wiggliness' penalty to the objective function used to estimate {β}:

$$\sum_i \left(y_i - \sum_j \beta_j X_{ij}\right)^2 + \lambda \cdot \int \left[f''(x)\right]^2 dx$$

- The penalty term removes the pressure to choose just the right set of knots.
- In case you're skeptical, let's try it.

Back to Our Example

- With "manual" spline regression we were judicious in our placement of knots.
- With GAM, we can err on the side of liberalism.
- **A 30-knot GAM slightly outperforms both a 10-knot GAM and our 5-knot spline regression.**
- **A 100-knot GAM is virtually indistinguishable from the 30-knot GAM!**
  - Run time is the primary disadvantage of choosing too many knots.



Generalized Additive Models for Geo-Spatial Analysis

Background – Territorial Ratemaking

- Common techniques for reflecting geography in insurance models:
    - Credibility models
    - Adding geo-demographic, crime, weather, traffic … variables to models
    - Spatial smoothing concepts

- Generalized Additive Models are a practical way to incorporate spatial smoothing in one's model.

- Some advantages:
    - Familiar paradigm: GAM is a generalization of GLM
    - Latitude and longitude can be used as model inputs
    - Lat/long can be incorporated alongside demographic variables
    - Use of offsets enables "modular" approach

Standard references:

- *Generalized Additive Models* by Hastie and Tibshirani (not tied to spline regression)
- *Generalized Additive Models* by Simon Wood (**paradigm followed here**)

California House Value Data

- One record per California block group.
- Target:
  - median house value
- Predictors:
  - Median income
  - Median house age
  - Average # bedrooms
  - Latitude
  - Longitude
- Let's fit a traditional GLM model on the first 3 predictors, and then bring in lat/long.

```
> ca.houses[1:10,]
    value income age  bedrooms   lat    long
1  452600 8.3252  41 0.4006211 37.88 -122.23
2  358500 8.3014  21 0.4606414 37.86 -122.22
3  352100 7.2574  52 0.3830645 37.85 -122.24
4  341300 5.6431  52 0.4211470 37.85 -122.25
5  342200 3.8462  52 0.4955752 37.85 -122.25
6  269700 4.0368  52 0.5157385 37.85 -122.25
7  299200 3.6591  52 0.4469835 37.84 -122.25
8  241400 3.1200  52 0.5937770 37.84 -122.25
9  226700 2.0804  42 0.5514096 37.84 -122.26
10 261100 3.6912  52 0.4558349 37.84 -122.25
>
>
> round(cor(ca.houses),2)
         value income   age bedrooms   lat  long
value     1.00   0.70  0.11     0.20 -0.14 -0.05
income    0.70   1.00 -0.12    -0.07 -0.08 -0.02
age       0.11  -0.12  1.00    -0.03  0.01 -0.11
bedrooms  0.20  -0.07 -0.03     1.00  0.16 -0.13
lat      -0.14  -0.08  0.01     0.16  1.00 -0.92
long     -0.05  -0.02 -0.11    -0.13 -0.92  1.00
```

The GAM is Afoot

Methodology:

1. Fit Gamma **GLM** to model house value as a linear combination of:
   - Income
   - Age
   - # Bedrooms

$$\log(VALUE) = \alpha + \beta_1 INCOME + \beta_2 AGE + \beta_3 ROOMS$$

2. Calculate the linear predictor for each data point: $\eta \equiv \beta \cdot X$

$$\eta \equiv \hat{\alpha} + \hat{\beta}_1 INCOME + \hat{\beta}_2 AGE + \hat{\beta}_3 ROOMS$$

3. Fit a Gamma **GAM** on *f*(lat,long) using η as an offset.

$$\log(VALUE) = \eta + f(lat, long)$$

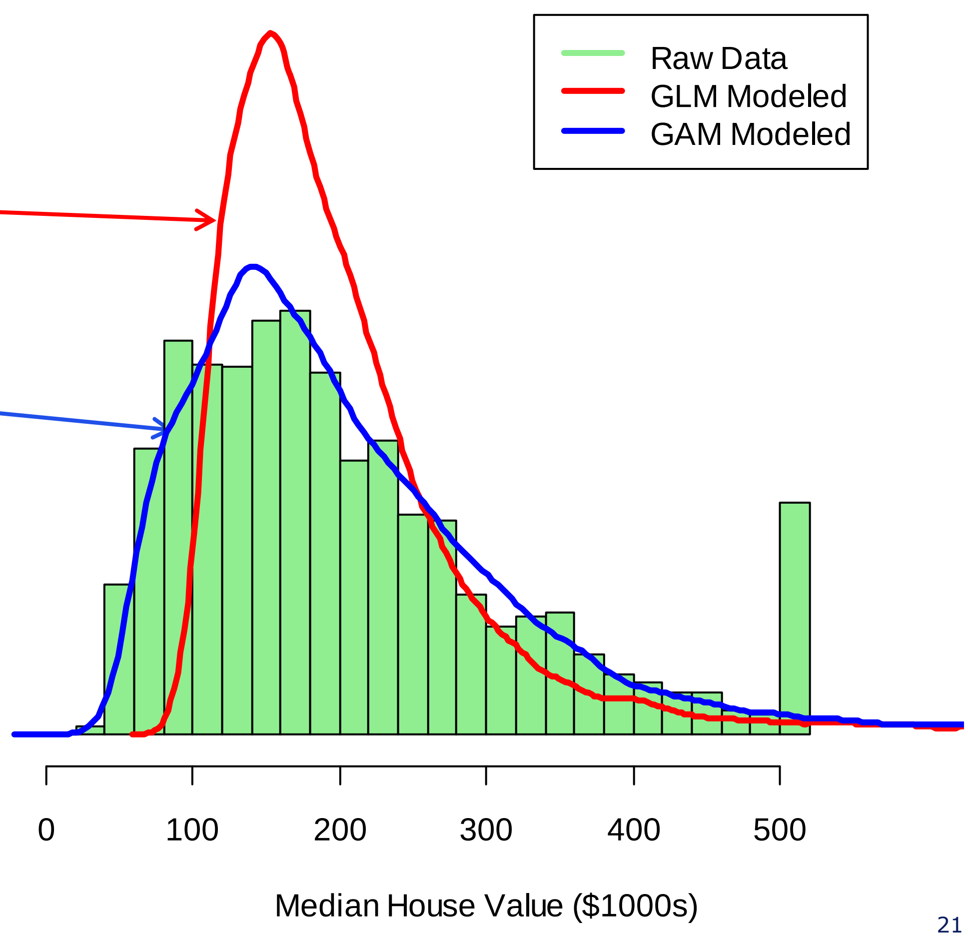
   - Note: For this illustration, tensor product basis functions with 400 knots were used.

Score Distributions

$$\log(VALUE) = \alpha + \beta_1 INCOME + \beta_2 AGE + \beta_3 ROOMS + f(lat, long)$$

- The 3-factor GLM doesn't come close to capturing all of the variation in house values.
- Adding *f*(location) helps.

California Median House Values (Block Group-Level)

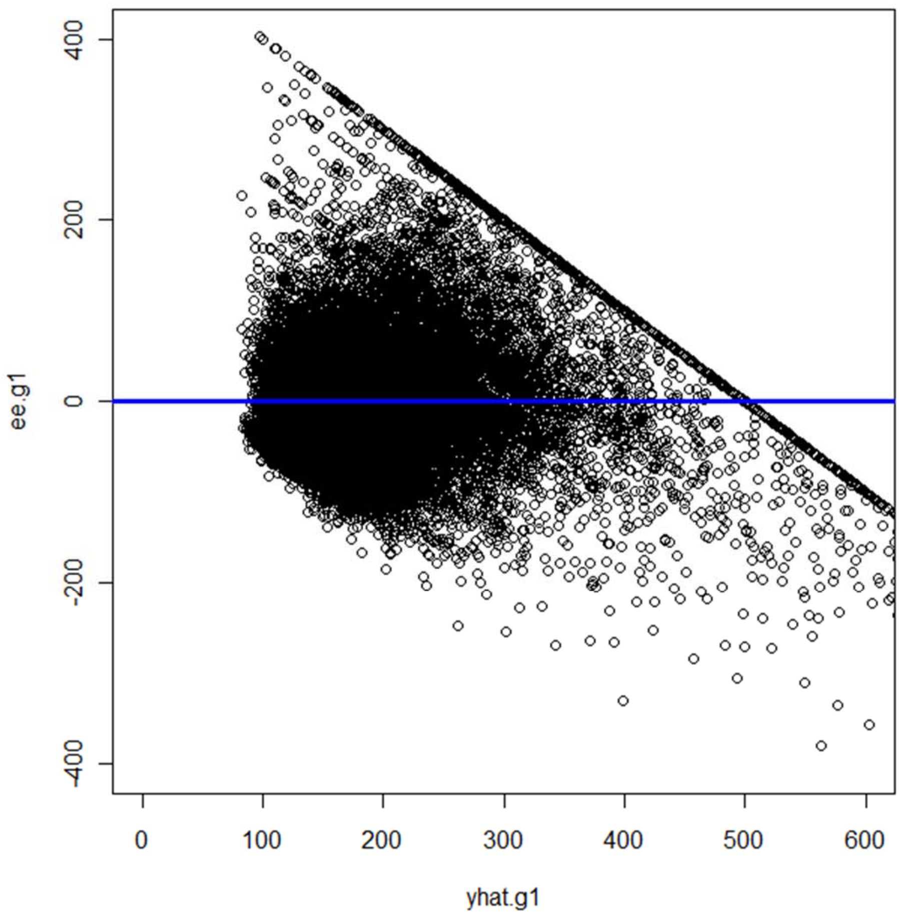


Error Diagnostics

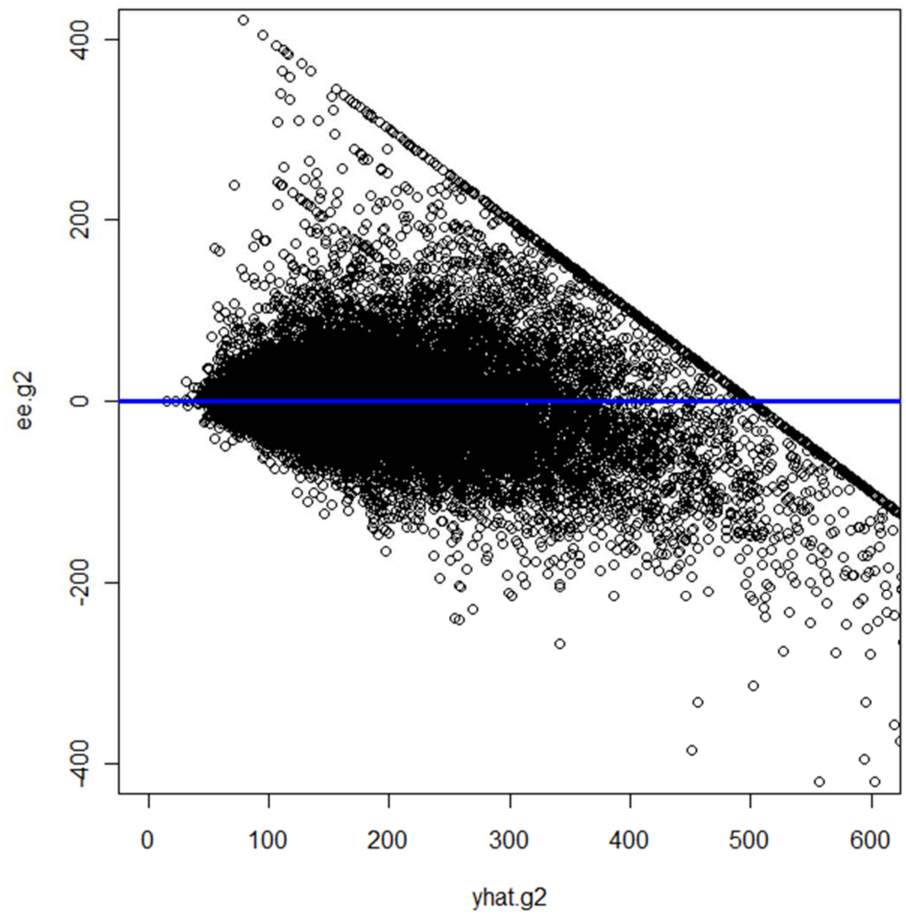
- The GAM model clearly explains more of the variation in house values.
  - R^2 GLM: 0.54
  - R^2 GAM: 0.67

Note: Raw Data capped at $500K – accounts for unusual residual pattern.

GLM Model (g1)

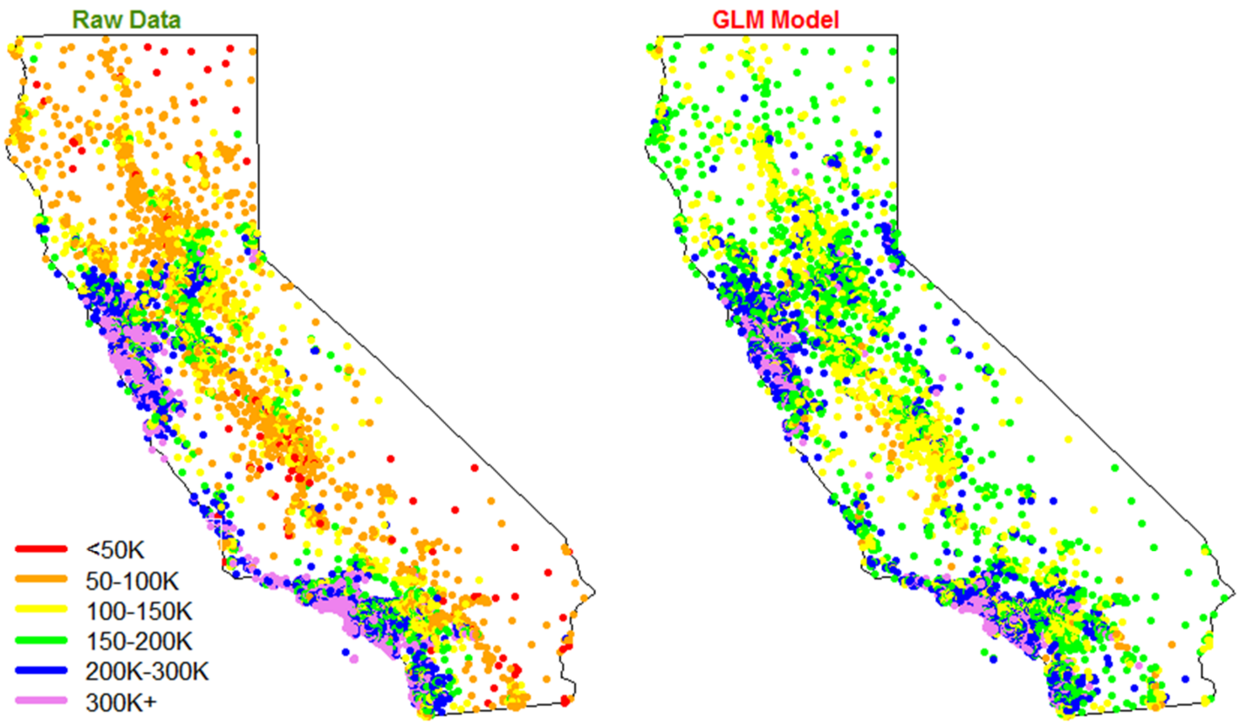


GAM Model (g2)



Geo-Spatial Diagnostics of the GLM Model

- The 3-factor GLM gets things directionally right:
  - Inland house values are lower than coastal house values
  - High values clustered around the major cities

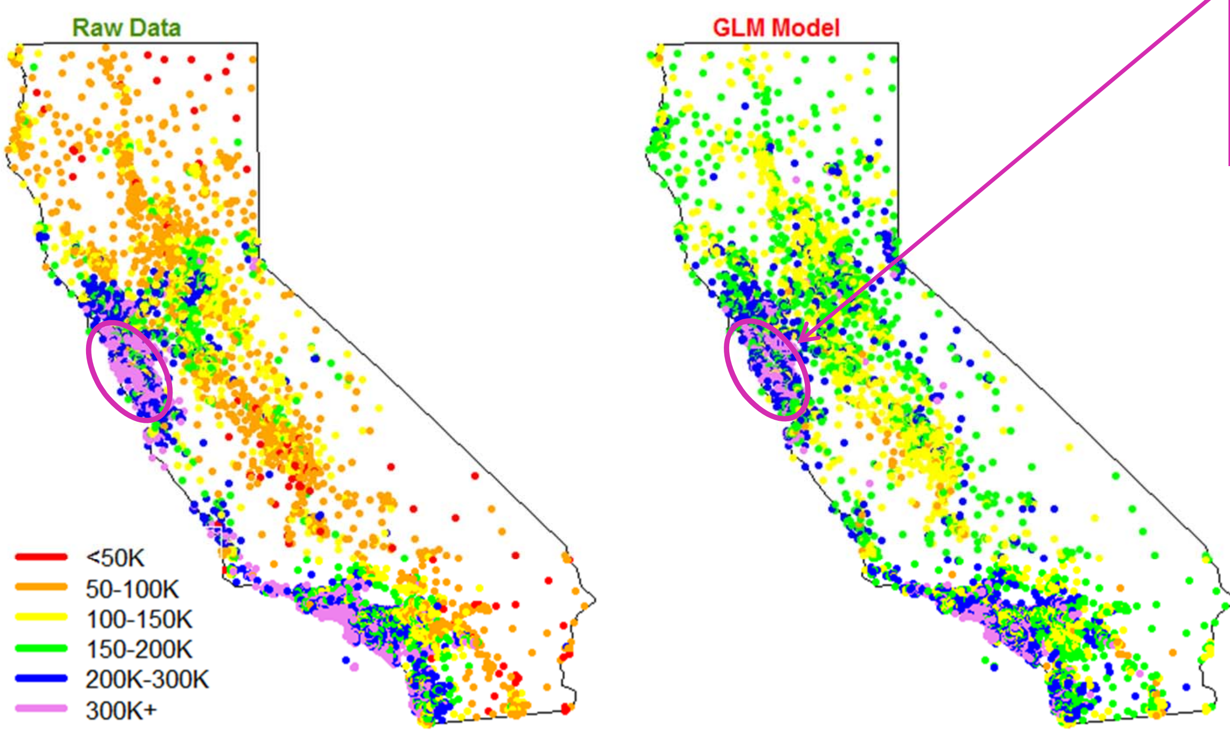


$$\log(VALUE) = \alpha + \beta_1 INCOME + \beta_2 AGE + \beta_3 ROOMS$$

Geo-Spatial Diagnostics of the GLM Model

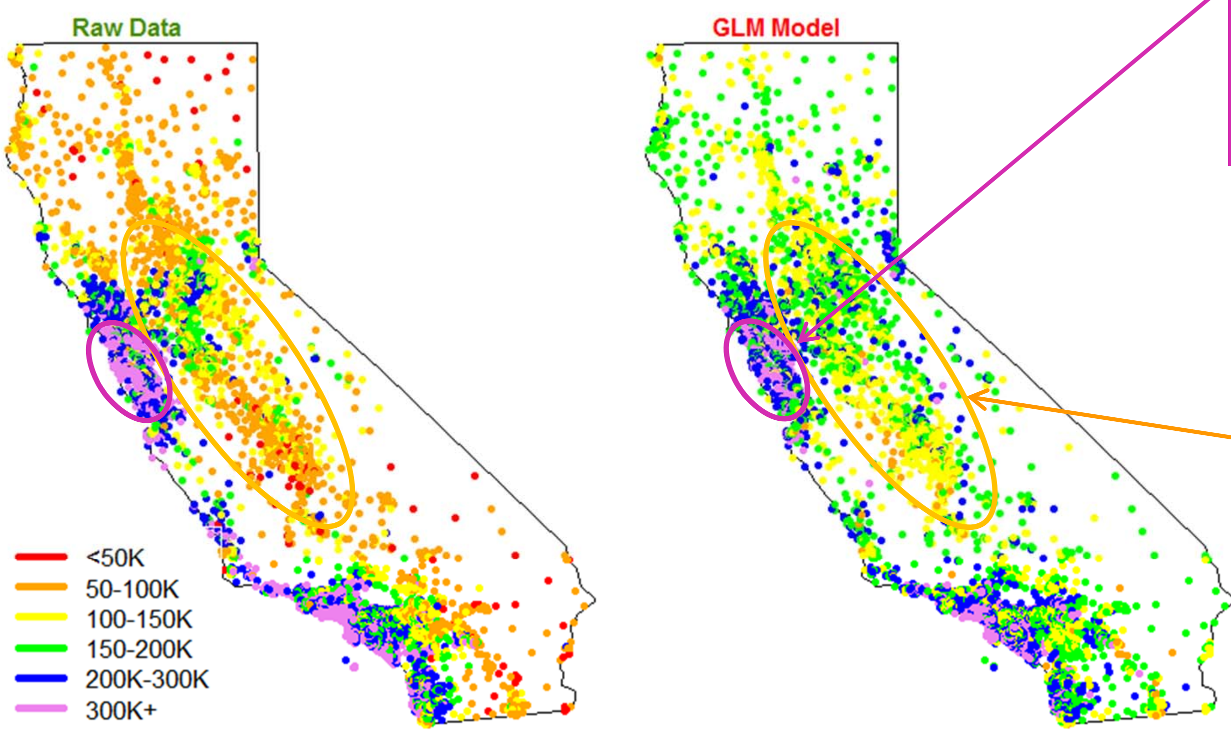
- But the GLM model generally:
  - Over-estimates house values in the central valley
  - Under-estimates house values in along the coast

The 3-factor GLM model generally <u>under</u>-estimates the coastal enclaves of expensive houses.



Geo-Spatial Diagnostics of the GLM Model

- But the GLM model generally:
  - Over-estimates house values in the central valley
  - Under-estimates house values in along the coast

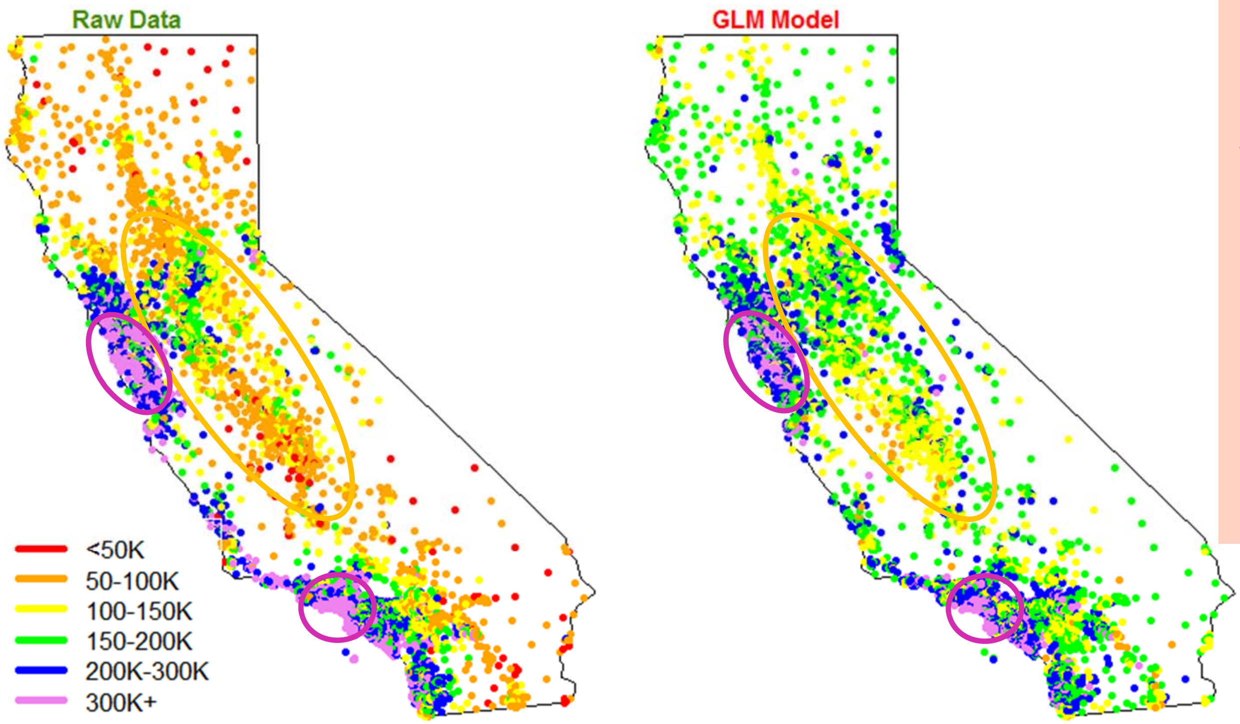


The 3-factor GLM model generally <u>under</u>-estimates the coastal enclaves of expensive houses.

At the same time, it <u>over</u>-estimates house values in the central valley

Location, Location, Location

- Implication: "Location matters."
- The GLM model shoves geo-spatial variation into the error term.



All else equal, houses in coastal / urban areas are worth more than houses in rural / inland areas.

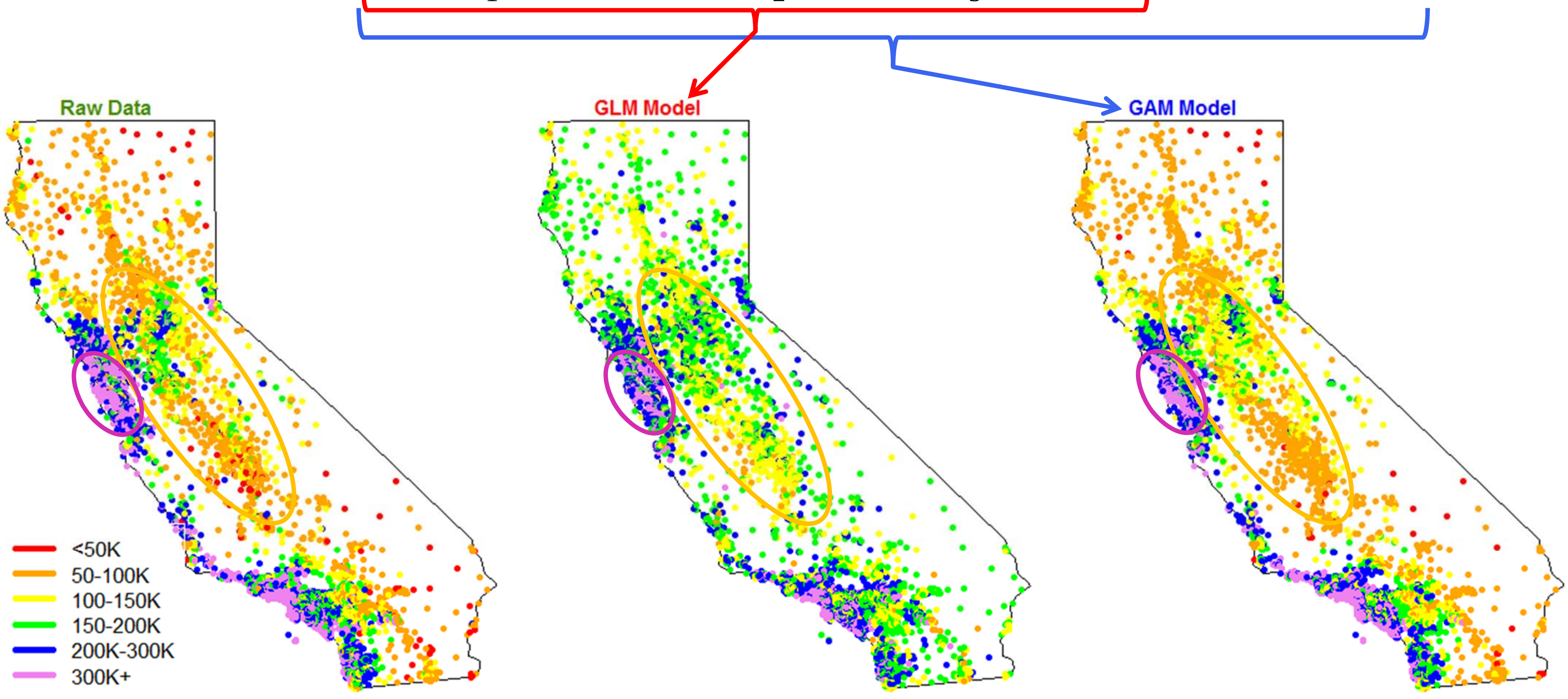
Adding further demographic predictors will help, but not eliminate the need to include location in the model.

-- miles from the coast
-- population density
-- education
-- neighborhood amenities
-- ...

GAM Diagnostics

The GAM model is still not perfect, but a big improvement over the 3-factor GLM model.

$$\log(VALUE) = \alpha + \beta_1 INCOME + \beta_2 AGE + \beta_3 ROOMS + f(lat, long)$$



Further improvements could result from superimposing one or more local GAM models built for specific metropolitan areas.