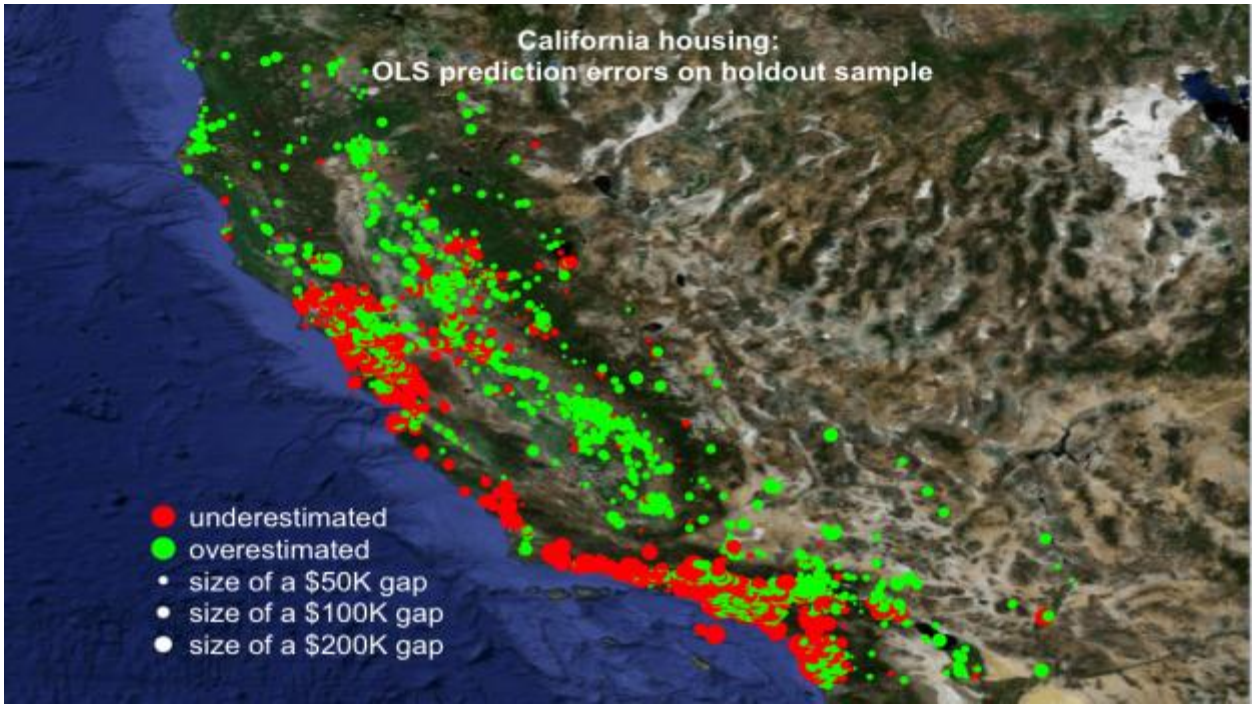
Location matters.
3 techniques to incorporate geo-spatial effects in one's predictive model

Xavier Conort
xavier.conort@gear-analytics.com

Motivation

- **Location matters!**
  - Observed value at one location is influenced by the observed values at other locations in a geographic area.
  - True for real estate, species distribution, insurance…
- Classical approach consists of adding different proxy variables (geo-demographic, crime, weather, traffic…) in predictive models. Yet it is usually not enough to capture all the geographical information available
- Idea: use of modern techniques to **incorporate latitude / longitude information as model inputs**
  - Generalized Additive Models (GAMs), Spatial Simultaneous Autoregressive Error Models (SARs) and Boosted Regression Trees (BRTs/GBMs) are 3 practical tools to boost models accuracy

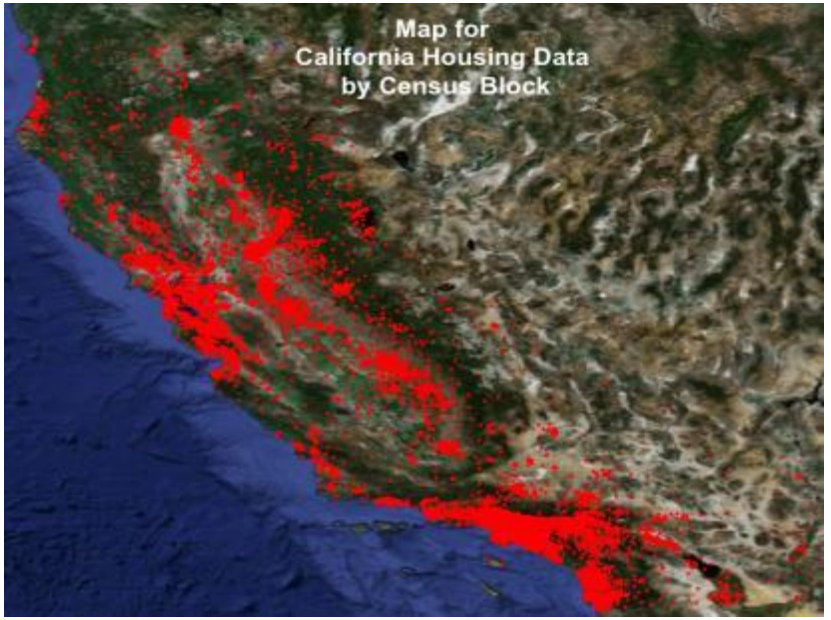
The California Housing example



Use of proxy variables (median income, housing density, average occupation in each house) is not enough and leads to an underestimation in coastal areas and overestimation in inland areas.

The California Housing Dataset

- The data set was first introduced by Pace and Barry (1997) and is available @
  - http://www.liaad.up.pt/~ltorgo/Regression/cal_housing.html
- It consists of aggregated data from each of 20,460 neighbourhoods (1990 census block groups) in California
- The response variable Y is the median house value in each neighbourhood



- There are a total of eight predictors, all numeric.
  - demographics variables (median income, housing density and average occupancy in each house).
  - the location of each neighbourhood (longitude and latitude),
  - and properties of the houses in the neighbourhood (average number of rooms and bedrooms).

Geo-spatial modelling: 3 different techniques as candidate

- **SARs are specifically designed to account for geographical effects.**
- Sometimes presented as an extension of time series in 2D.
- Use of neighbourhoods to specify the spatial correlation structure allow them to work on large datasets

- **BRTs belong to the machine learning field**.
- Not specifically designed to account for spatial autocorrelation but BRTs' ability to learn local interactions should make them a good candidate

- **Generalized Additive Models (GAMs) are GLMs which can contain smooth functions of covariates**. Use of a smooth function of Latitude-Longitude as a predictor can be a practical way to incorporate spatial smoothing

How do SARs work ?

Approach 1: specify correlation structure

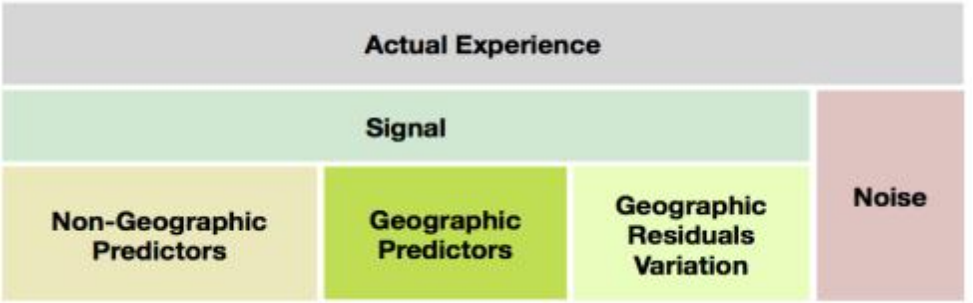
- SARs assume that the response at each location i is a function not only of the explanatory variables at i, but of the values of the response at neighbouring locations j as well
- The neighbourhood relationship is formally expressed in a n × n matrix of spatial weights (W), with elements (w_{ij}) representing a measure of the connection between locations i and j
- SAR Error Model

  $\underline{Y = X\beta + u}$

  $\underline{with \quad u = \lambda W u + \varepsilon}$

  λ= *Spatial Lag Coefficient*

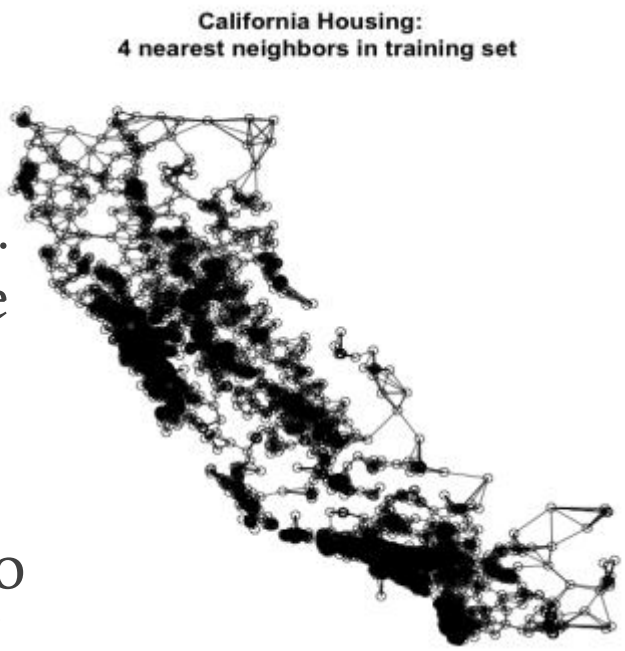
  ε= *normal error*



See Sengupta's presentation in 2010 CAS RPM Seminar

The neighbourhood structure

- The neighbourhood can be identified by
  - a fixed number of nearest neighbors,
  - or by Euclidean or great circle distance (e.g. the distance along Earth's surface) to define cells within or outside a respective neighbourhood
- The neighbours can further be weighted to give closer neighbours higher weights and more distant neighbours lower weights.



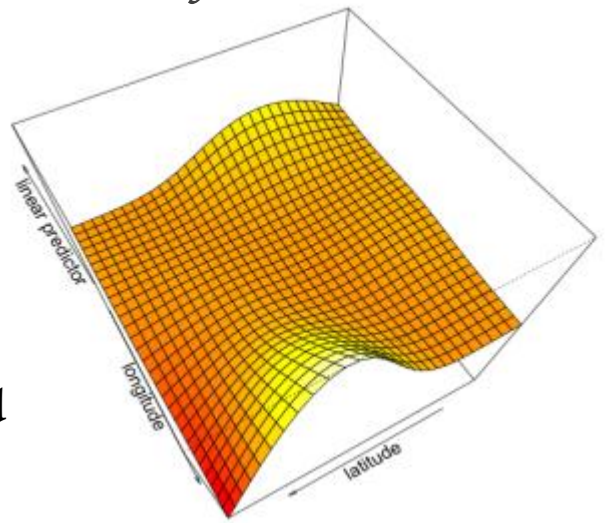
How do GAMs work ?

Approach 2: smooth geographical trends

GAMs **use the basic ideas of Generalized Linear Models**

- $g(\mu) \equiv g(E[Y]) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_N X_N$
- $Y \mid \{X\} \sim$ exponential family

- While in GLMs $g(\mu)$ is a linear combination of predictors, in GAMs the linear predictor can also contain one or more *smooth functions* of covariates
  - $g(\mu) = \beta \cdot \boldsymbol{X} + f_1(X_1) + f_2(X_2) + f_3(X_3, X_4) + \ldots$
  - To represent the functions f, use of cubic splines is common
  - To avoid over-fitting, a penalized Maximum Likelihood (ML) is minimized instead of the ML in GLMs.
  - The optimal penalty parameter is automatically obtained via cross-validation



See Guszcza*'s presentation in 2010 CAS RPM Seminar*

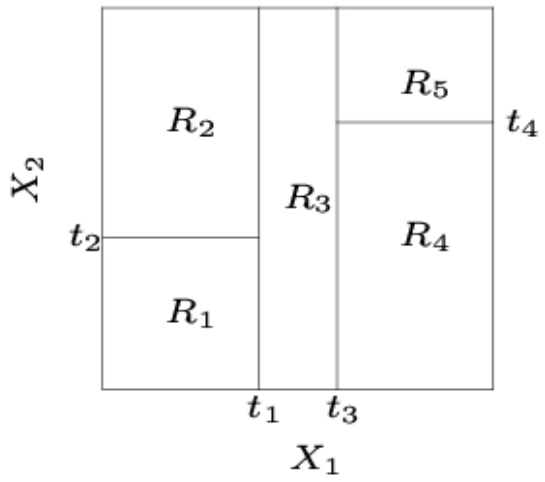
How do BRTs work ?

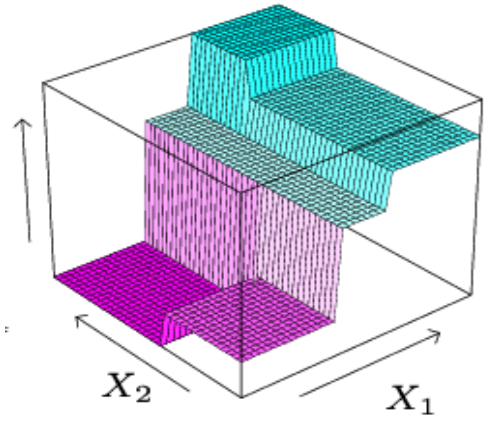
Approach 3: let the machine do if for you!

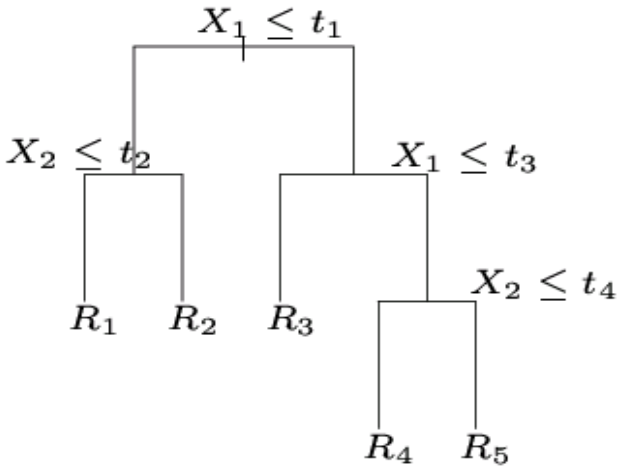
- BRTs (also called Gradient Boosting Machine) use **boosting and decision trees techniques:**
  - The boosting algorithm learns step by step slowly and gradually increases emphasis on poorly modelled observations. It minimizes a loss function (the deviance, as in GLMs) by adding, at each step, a new simple tree whose focus is only on the residuals
  - The contributions of each tree are shrunk by setting a learning rate very small (and < 1) to give more stable fitted values for the final model
  - To further improve predictive performance, the process uses random subsets of data to fit each new tree (bagging).

Boosted Regression trees are good to detect interactions

… as they are based on regression trees which partition the feature space into a set of rectangles and then produce a multitude of local interactions.
This totally makes sense whilst modelling geographical effect.







Other nice properties of BRTs

- BRTs can be fitted to a **variety of response types** (Gaussian, Poisson, Binomial)
- BRTs best fit (interactions included) is **automatically detected by the machine**
- BRTs learn **non-linear functions** without the need to specify them
- BRT outputs have some GLM flavour and **provide insight** on the relationship between the response and the predictors
- BRTs **avoid doing much data cleaning** because of their
  - ability to accommodate missing values
  - immunity to monotone transformations of predictors, extreme outliers and irrelevant predictors

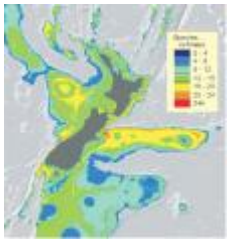
BRTs examples links

- Orange's churn, up-, and cross-sell at 2009 KDD Cup
  - http://jmlr.csail.mit.edu/proceedings/papers/v7/miller09/miller09.pdf

- Yahoo Learning to Rank Challenge
  - http://jmlr.csail.mit.edu/proceedings/papers/v14/chapelle11a/chapelle11a.pdf

- Patients most likely to be admitted to hospital - Health Heritage Prize
  - Only available to Kaggle's members

- Fraud detection in
  - http://www.data-mines.com/Resources/Papers/Fraud%20Comparison.pdf



- Fish species richness
  - http://www.stanford.edu/~hastie/Papers/leathwick%20et%20al%202006%20MEPS%20.pdf

- BRTs in those papers are sometimes called a different way: Gradient Boosting Machine, TreeNet…. But they are the same models.

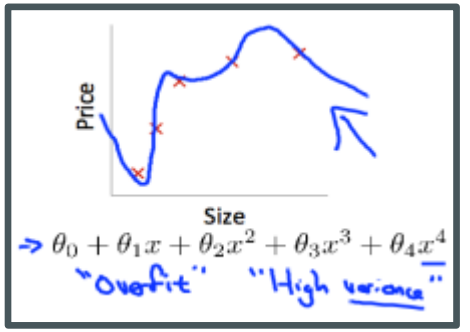
Software used

Entire analysis (including all graphics) in this presentation is done in 

- GAMs: Wood's package (**mgcv**).
- SARs: **spdep**, a spatial econometrics package developped by Bivand et al. We, however, wrote our own code for predictions.
- BRTs: Ridgeway's package (**gbm**)
- We plotted maps with Loecher's package (**RgoogleMaps**)which take Google's map as a background image to overlay plots within R

Strategy to assess models performance

- Many models can be made to perform well on their training data, the danger is that they overfit to specific features of that data that lack applicability in a wider sample, degrading model performance when predicting to new datasets.

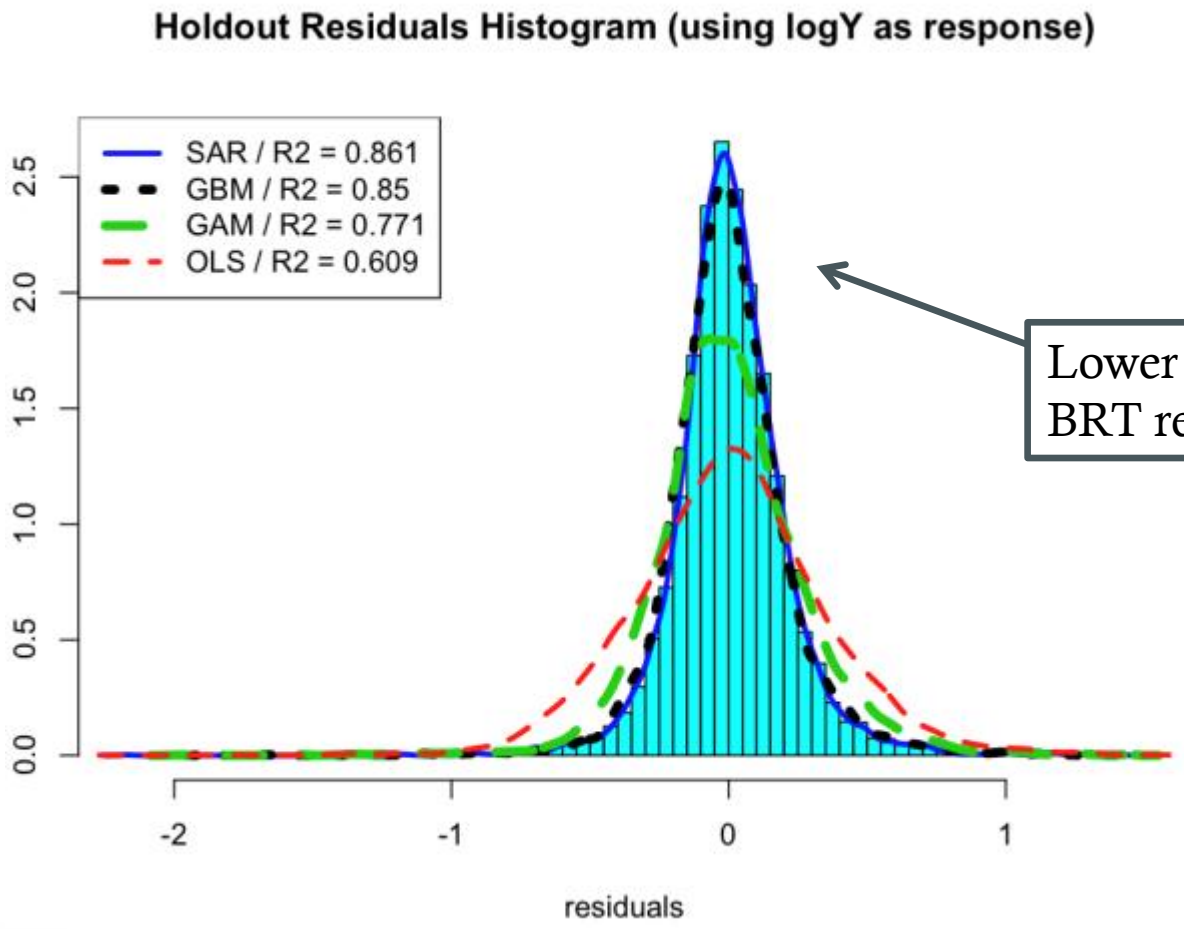


- **Best practice consists in assessing model predictive performance using independent data** (cross-validation)
  - **Partitioning** the data into separate training and testing subsets
  - **k-fold** cross-validation **when data is scarce**

1	2	3	4	5
Train	Train	Test	Train	Train

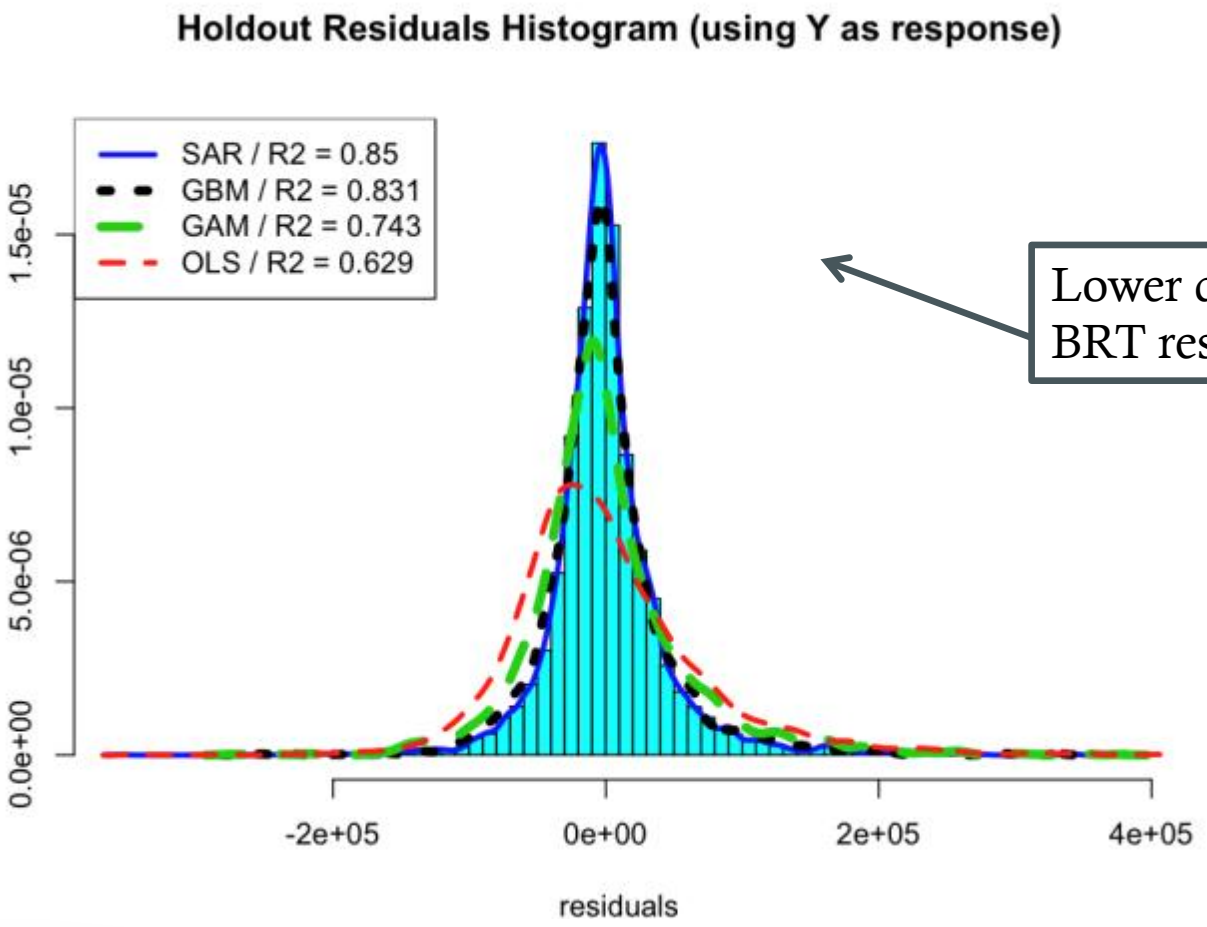
- Here, we chose to partition the data in a training set (70% of examples) and a testing set (30%) and then reported
  - errors maps and errors distributions,
  - and the squared multiple correlation coefficient R^2 (the proportion of variance in Y that can be accounted for by our model) measured on the testing set as previously chosen in previous studies (with both logY and Y as the response).

Residuals histograms (with logY as the response)



Lower dispersion for SAR and BRT residual distributions

Residuals histograms (with Y as the response)

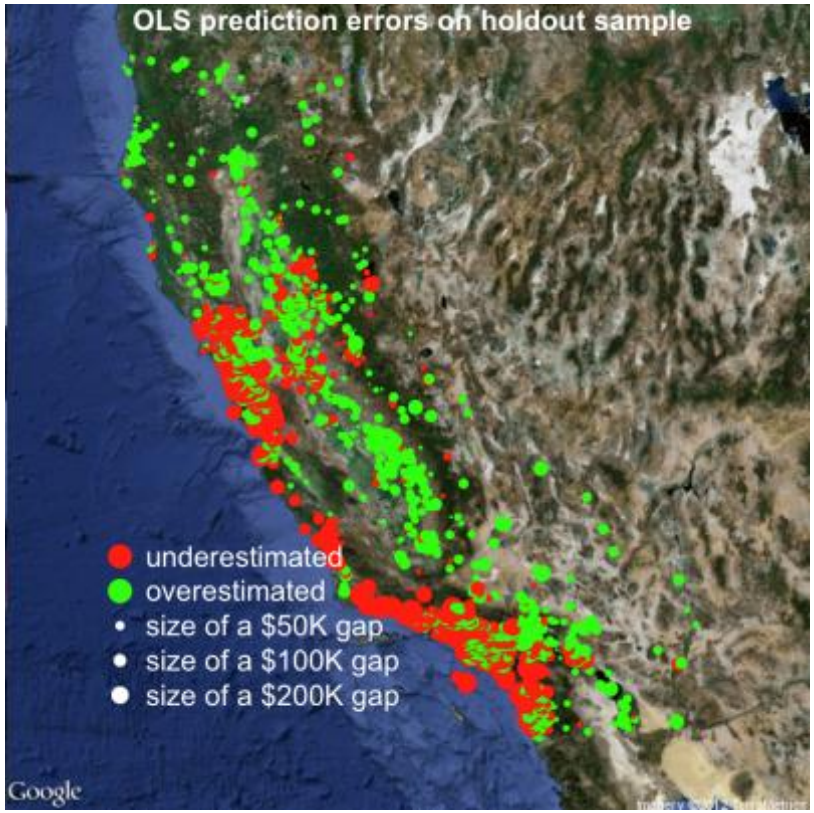


Lower dispersion for SAR and BRT residual distributions

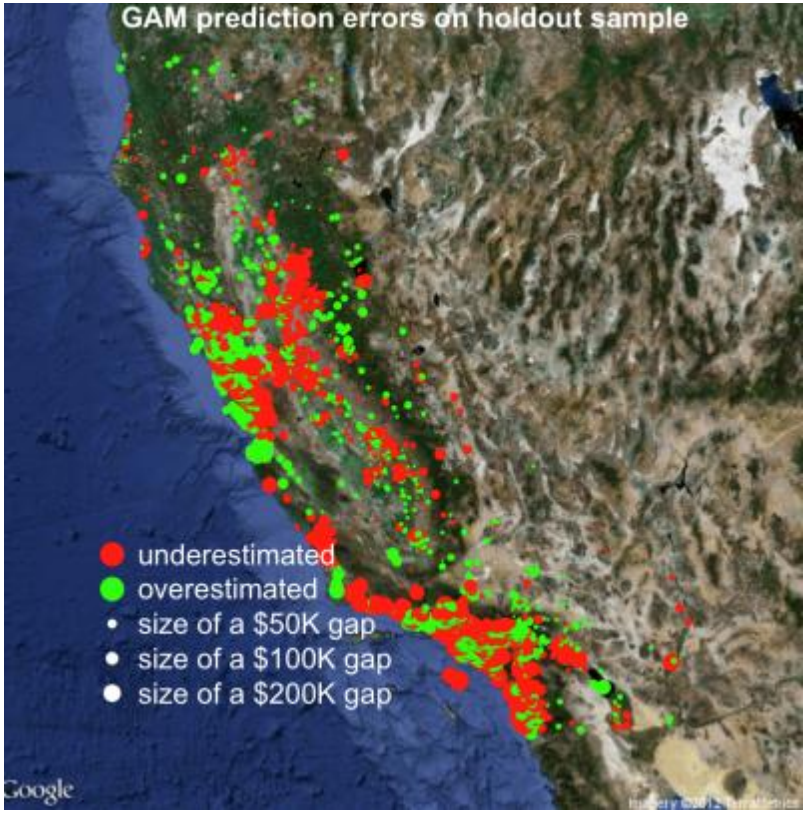
To predict Y, we fitted a GAM to model the relationship between E(Yi) and E(log(Yi)).
Works better than assuming that E(Yi)=exp(E(log(Yi)+sigma^2/2) with sigma constant.

Maps outputs (OLS vs GAMs)

Ordinary Least Squares

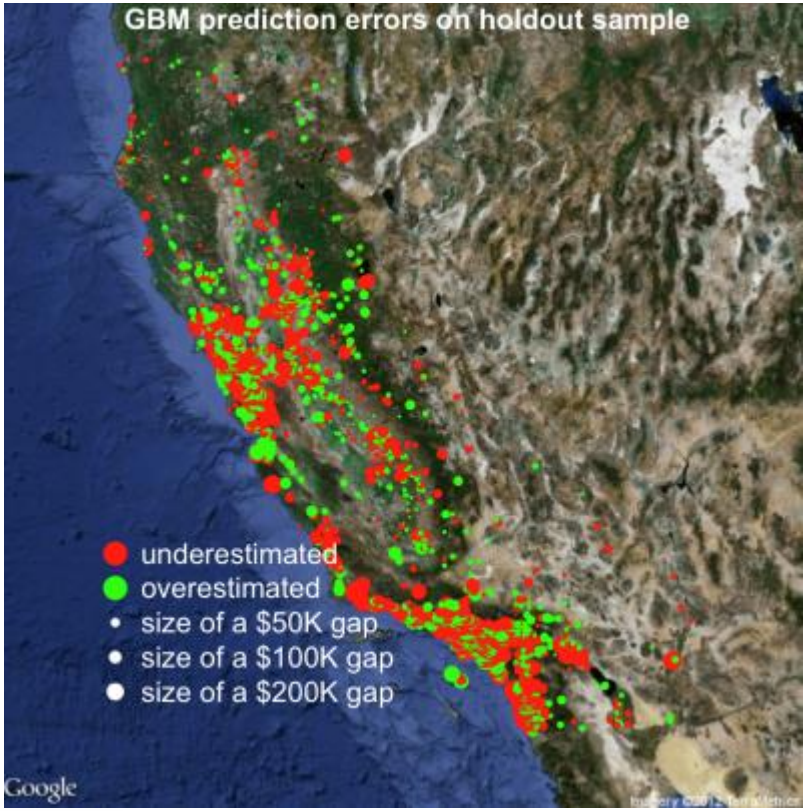


Generalized Additive Models

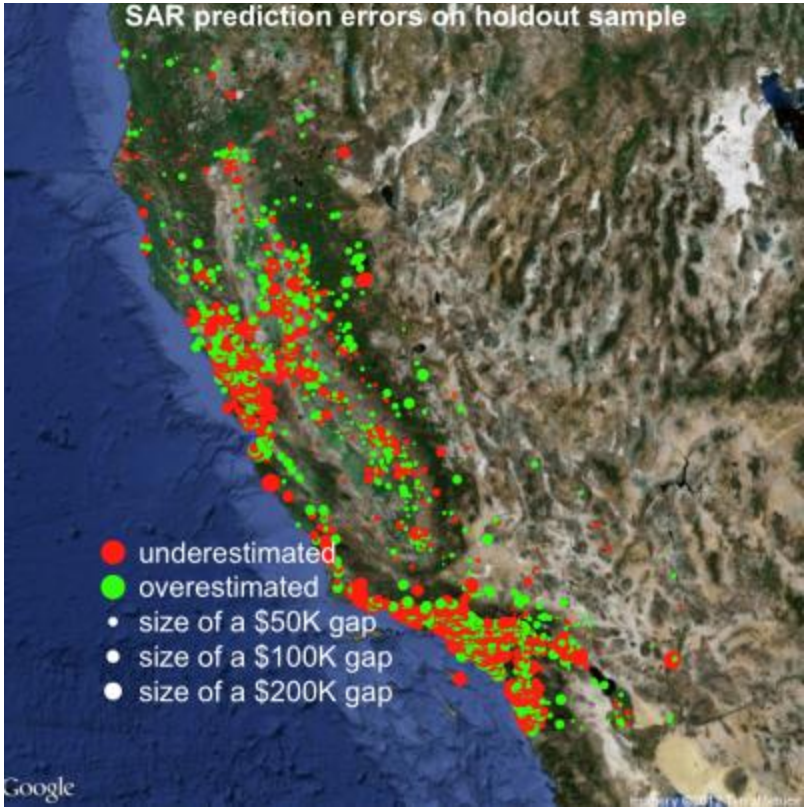


Maps outputs (BRTs vs SAR)

BRTs



SAR and 4 nearest neighbors



Strategies to improve predictive accuracy

- Fine tune the number of nearest neighbours for SAR
- Find a more appropriate transformation for the response than the logY suggested by Pace & Barry in their 1997 paper
- Split analysis of non censored/censored house values
- Fit SARs on GAM residuals
- Fit SARs on BRT residuals
- Fit BRTs on SAR residuals
- If you are lazy, try BRTs*0.5 + SAR*0.5 !!!

Ideas to boost your modelling in insurance

- When only predictive accuracy matters
  - Use BRTs to get high predictive power with low modelling effort
  - If you want even better accuracy, combine BRTs with Random Forest (another off-the-shell ML technique)
  - in presence of spatial correlation, try SARs+BRTs
- When you want to keep control on the model structure such as in pricing
  - Fit GLMs / GAMs
    - fit BRTs for an easy benchmark and highlight useful interactions
    - If you have variables with many categories (such as districts) use GLMMs to get credibility estimates
  - In presence of spatial correlation, try GAMs or SARs / BRTs on GLMs residuals

To know more on Machine Learning, go to Stanford's online course ml-class.org by Andrew Ng