A Change Point Model for Housing Values

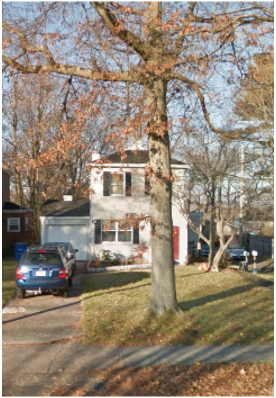
Xiaofei (Susan) Wang

Lecturer/Research Scholar
Department of Statistics
Yale University

JSM 2016

Joint work with Jay Emerson.

Two Houses in New Haven

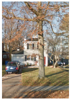


638 Prospect St.

308 Shelton Ave.

Photos from Google Street View.

Two Houses in New Haven



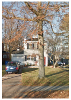
638 Prospect St.

308 Shelton Ave.

Property	638 Prospect	308 Shelton
lot size	0.11 acres	0.12 acres
living area	1440 sq.ft.	3295 sq.ft.
# bedrooms	2	4
# bathrooms	1	2

Two Houses in New Haven

Actual assessed values (2011):



638 Prospect St.

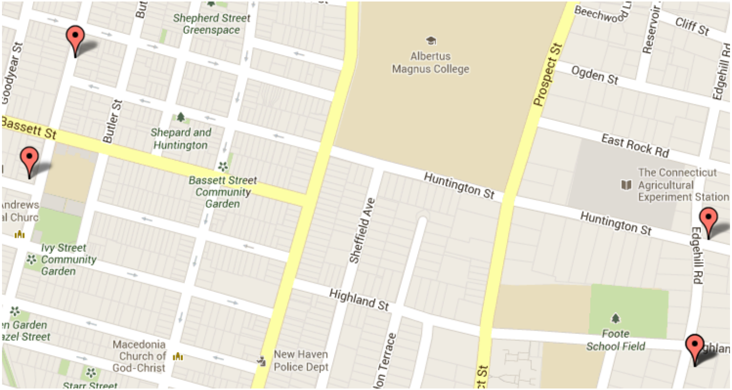
308 Shelton Ave.

Property	638 Prospect	308 Shelton
lot size	0.11 acres	0.12 acres
living area	1440 sq.ft.	3295 sq.ft.
# bedrooms	2	4
# bathrooms	1	2
assessed value	$219,100	$191,310

More Houses

$n = 244$ houses in New Haven, CT

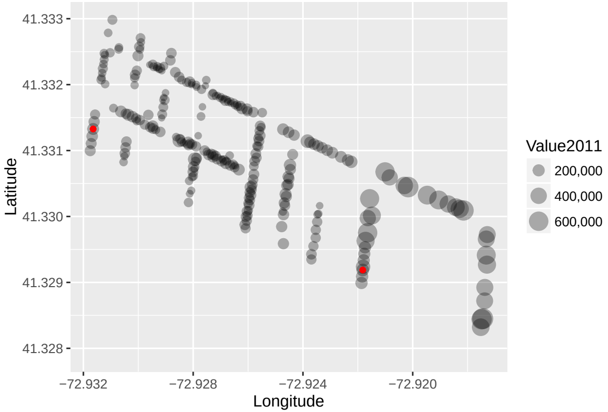
Goal: To model 2011 assessed values using available information.



More Houses

$n = 244$ houses in New Haven, CT

Goal: To model 2011 assessed values using available information.



Red circles indicate locations of two houses previously considered.

A Naive Approach

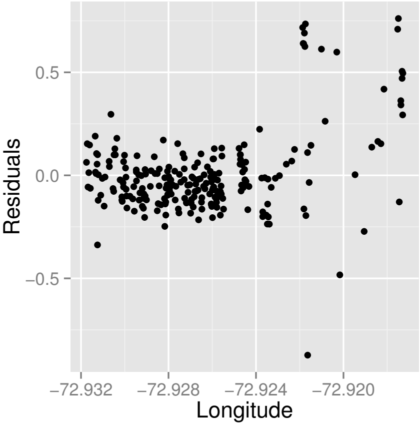
Model 1: Linear Regression

```
lm(log(assessed value) ~ lot size + sqrt(living area) +
                          # bedrooms)
```

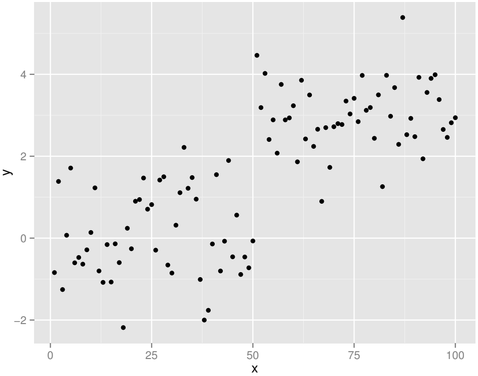
	Estimate	Std. Error	t value	p value
(Intercept)	10.5	0.067	155.7	$<$2e-16
$\sqrt{\text{living area}}$	0.028	0.0022	12.89	$<$2e-16
beds	-0.036	0.012	-3.13	0.0020
size	1.53	0.11	14.15	$<$2e-16

Residual standard error: 0.200 on 240 degrees of freedom
Multiple R-squared: 0.82

Model 1 Residuals

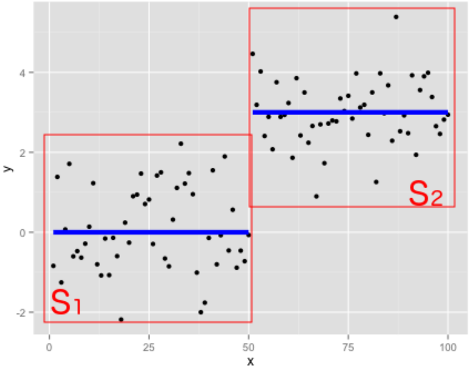


What are change points?



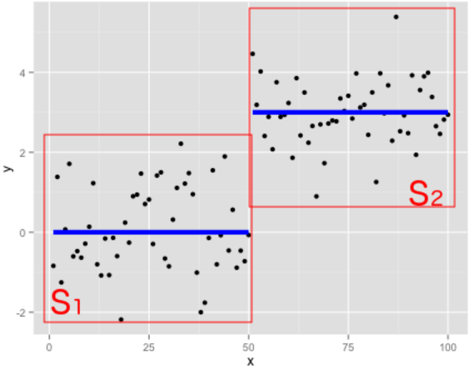
- Observations: $y_1, \ldots, y_{100}$

What are change points?



- Observations: $y_1, \dots, y_{100}$
- Partition: $\rho = (S_1, S_2)$
  $S_1 = \{1, \dots, 50\}$
  $S_2 = \{51, \dots, 100\}$
- $y_{i:i \in S_1} \sim N(0, 1)$
  $y_{i:i \in S_2} \sim N(3, 1)$
- Change point at location 50.

What are change points?



- Observations: $y_1, \ldots, y_{100}$
- Partition: $\rho = (S_1, S_2)$
  $S_1 = \{1, \ldots, 50\}$
  $S_2 = \{51, \ldots, 100\}$
- $y_{i:i \in S_1} \sim N(0, 1)$
  $y_{i:i \in S_2} \sim N(3, 1)$
- Change point at location 50.

Change points partition observations into blocks. Within each block, observations share a common distribution.

Related Work

- Barry and Hartigan (1993, 1994), Erdman and Emerson (2007, 2008) - univariate, Bayesian (`library(bcp)`)
- Bai and Perron (2003), Zeileis, et al. (2001) - regression (`library(strucchange)`)
- Muggeo (2003) - regression (`library(segmented)`)
- Olshen, et al. (2004) - univariate, using circular binary segmentation
- Fearnhead (2005), Loschi, et al. (2010) - regression, Bayesian
- Killick and Eckley (2011) - univariate mean/variance (`library(changepoint)`)
- Matteson and James (2013) - multivariate (`library(ecp)`)
- ... and others

Classical Bayesian Change Point Analysis

Barry and Hartigan (1993): Univariate serial observations $y_i \sim N(\mu_i, \sigma^2)$

Classical Bayesian Change Point Analysis

Barry and Hartigan (1993): Univariate serial observations $y_i \sim N(\mu_i, \sigma^2)$

Partition $\rho = (S_1, \ldots, S_b)$

Prior on the partition:

$$\pi(\rho) \propto \int_0^{p_0} p^{b-1}(1-p)^{n-b}dp$$

Classical Bayesian Change Point Analysis

Barry and Hartigan (1993): Univariate serial observations $y_i \sim N(\mu_i, \sigma^2)$

Partition $\rho = (S_1, \ldots, S_b)$

Prior on the partition:

$$\pi(\rho) \propto \int_0^{p_0} p^{b-1}(1-p)^{n-b} dp$$

Likelihood:

$$y_{i:i \in S} | \theta_S, \sigma^2 \sim N(\theta_S, \sigma^2)$$

$$\theta_S | \mu_0, \sigma_0^2 \sim N\left(\mu_0, \frac{\sigma_0^2}{n_S}\right)$$

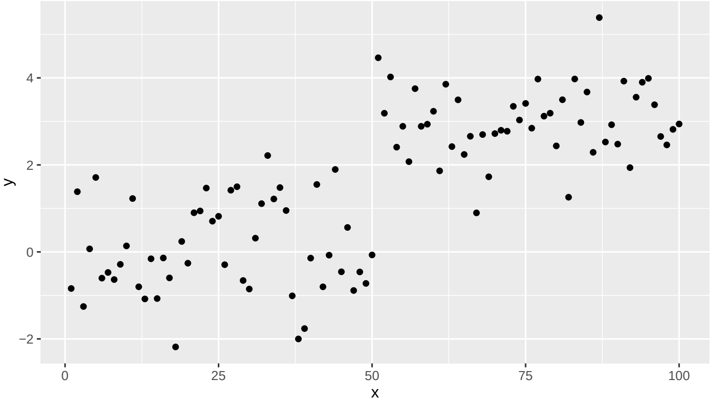
$$\mu_0 \sim U(-\infty, \infty)$$

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2} \quad \sigma^2 \in (0, \infty)$$

$$\pi(w) = \frac{1}{w_0'} \quad w \in (0, w_0')$$

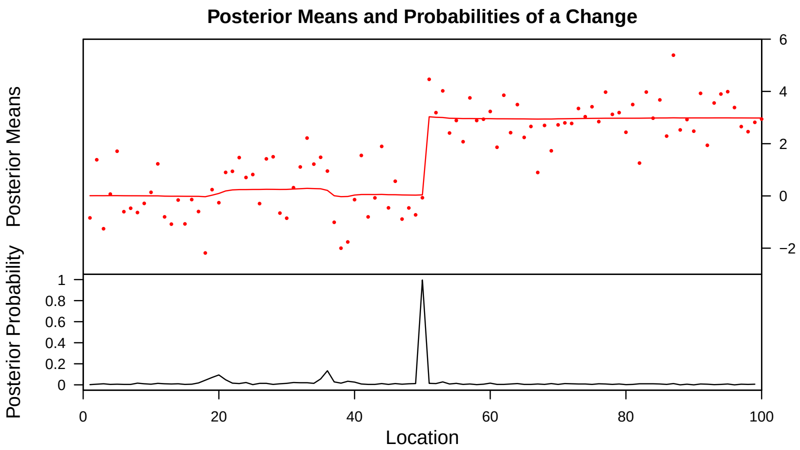
where $w = \sigma^2/(\sigma_0^2 + \sigma^2)$

Example



Example

Barry and Hartigan's approach implemented by Erdman and Emerson (2007, 2008) in `library(bcp)`:



Extensions to Linear Regression

Wang and Emerson (2016):

- Observations: $\{(\boldsymbol{x_i}, y_i)\}_{i=1}^n$
- y_i is scalar response at location i
- $\boldsymbol{x_i}$ is k-dimensional vector of predictors

Extensions to Linear Regression

Wang and Emerson (2016):

- Observations: $\{(\boldsymbol{x_i}, y_i)\}_{i=1}^{n}$
- y_i is scalar response at location i
- $\boldsymbol{x_i}$ is k-dimensional vector of predictors

Prior on the partition:

$$\pi(\rho) \propto \int_0^{p_0} p^{b-1}(1-p)^{n-b}dp$$

Likelihood:

$$y_{i:i\in S}|\boldsymbol{x}_i, \beta_S, \sigma^2 \sim N(\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}^{\boldsymbol{S}}\boldsymbol{\beta_S}, \sigma^2)$$

Prior on the intercept:

$$\beta_{S0}|\mu_0, \sigma_0^2 \sim N\left(\mu_0, \frac{\sigma_0^2}{n_S}\right)$$

Prior on other coefficients:

$$(\beta_{Sj}|\tau_S = 0) = 0 \quad \text{w.p. } 1$$

$$(\beta_{Sj}|\tau_S = 1, \sigma_j^2) \sim N\left(0, \frac{\sigma_j^2}{\sum_{i\in S}(x_{ij} - \bar{x}_{.j}^S)^2}\right)$$

Extensions to Linear Regression

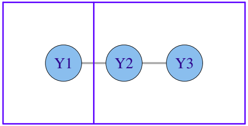
Other priors:

$$
\begin{aligned}
P(\tau_S = 0) &= \frac{d}{n_S + d}\mathbb{1}\{n_S \geq 2k\} + \mathbb{1}\{n_S < 2k\} \\
P(\tau_S = 1) &= \frac{n_S}{n_S + d}\mathbb{1}\{n_S \geq 2k\} \\
\mu_0 &\sim U(-\infty, \infty) \\
\pi(\sigma^2) &\propto \frac{1}{\sigma^2} \quad \sigma^2 \in (0, \infty) \\
\pi(w_j) &= \frac{1}{w_j} \quad w \in (0, w_j')
\end{aligned}
$$

where $w_j = \sigma^2/(\sigma_j^2 + \sigma^2)$

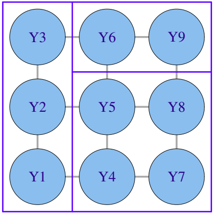
Extensions to Connected Graphs

Before: (serial data, residing on **path graph**)



$$f(\rho) \propto \int_0^{p_0} p^{b-1}(1-p)^{n-b} dp$$

Now: (**general graph**)

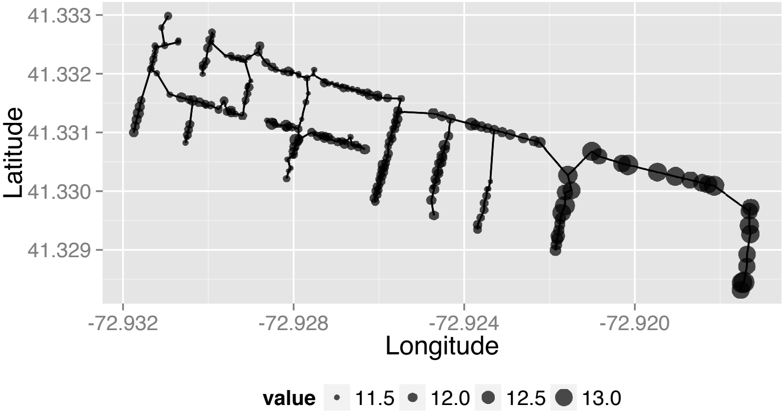


$$f(\rho) \propto \alpha^{l(\rho)}$$

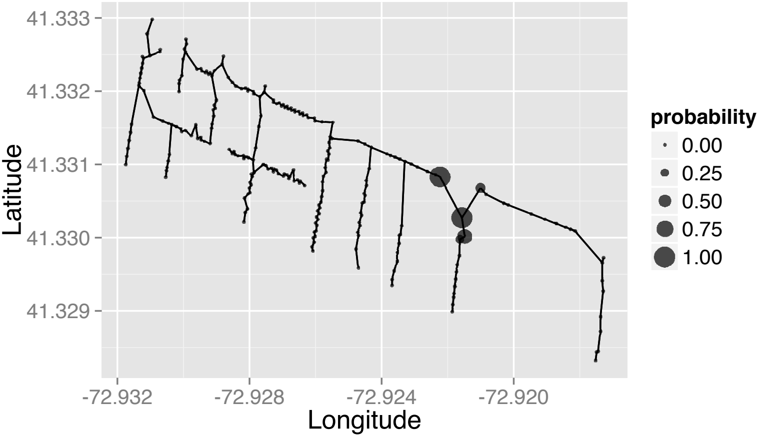
where $0 < \alpha < 1$ and $l(\rho)$ is a measure of boundary length

Housing Data Revisited

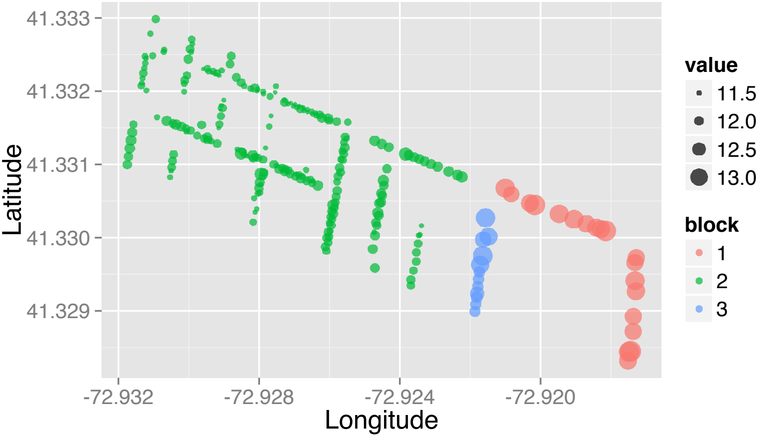
We need a graph structure to carry out change point analysis. Using a **minimum spanning tree**, we get:



BCP: Posterior Boundary Probabilities

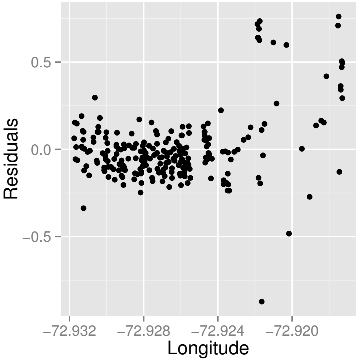


BCP: Modal Partition

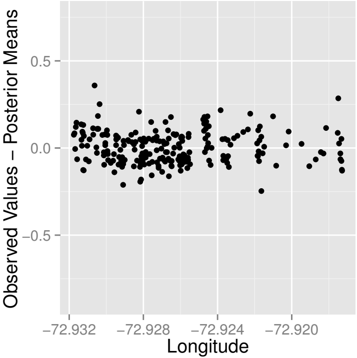


Model Comparison

Model 1: Regression
residual SE: 0.200

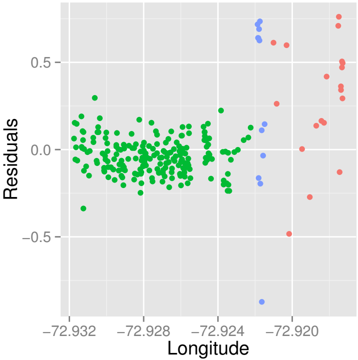


Model 2: BCP
SD("residuals"): 0.092

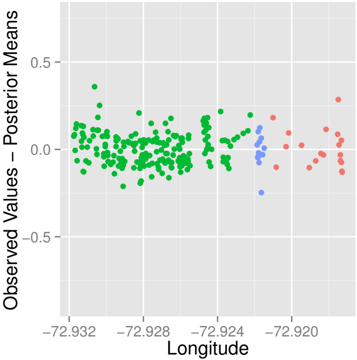


Model Comparison

Model 1: Regression
residual SE: 0.200



Model 2: BCP
SD(“residuals”): 0.092



Summary

Extension of Bayesian change point framework supports:

- regressions within blocks, if predicting variables are available
- data residing on nodes of a connected graph
- multivariate/univariate change point analysis

Thank You!

Preprint of article submitted to JASA:

- Under revision: Wang and Emerson (2016). "Bayesian Change Point Analysis of Linear Models on Graphs." `http://arxiv.org/abs/1509.00817`.

R package `bcp` is available on R CRAN:
`https://cran.r-project.org/web/packages/bcp/index.html`.