Scalable Gaussian Processes for Characterizing Multidimensional Change Surfaces

April 18, 2016

William Herlands

Committee: Daniel Neill, Alex Smola, Wilbert Van Panhuis

Chair: Dave Choi

Outline

1. Motivation
2. Gaussian process introduction
3. Change surface model
4. Analysis of measles in the United States

Complex Changes

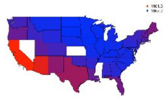
- In human systems changes are often complex
  - Policy interventions take time to trickle through government bureaucracy
  - Environmental hazards affect populations differentially

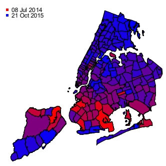
- **Simple changepoint models are not sufficiently expressive**

Why do we care?

- Understand past changes
  - Explore spatio-temporal heterogeneity
  - Model the rate of changes in different areas
- Enable more accurate or equitable policies

- Applications
  - Measles incidence in the U.S
  - Concerns about lead-tainted water in NYC





Our objectives

- Model complex changes in real world data
  - Multiple, flexible function regimes
  - Non-discrete changes
  - Non-monotonic changes
  - Heterogeneous changes over space, time, etc.

Gaussian processes for flexible functions

"Change surfaces" for complex changes

Gaussian Processes (GP)

- Non-parametric prior over smooth functions

$$f(x) \sim GP(m(x), k(x, x'))$$
$$m(x) = E[f(x)]$$
$$k(x, x') = \mathrm{cov}(f(x), f(x'))$$

- Covariance function is a kernel. Defines the covariance of function values

Gaussian Processes (GP)

- Any finite set of f(**x**) is Normally distributed

$$\left[f(x_1),\ldots, f(x_m)\right] \sim N(m(\mathbf{x}), K)$$

- Observation model

$$y(x) = f(x) + \varepsilon, \qquad \varepsilon \sim N(0, \sigma_\varepsilon)$$

- Marginal log likelihood optimization

$$\log p(\mathbf{y} \mid \theta) \propto -\log \left| K + \sigma_\varepsilon I \right| - \mathbf{y}^T (K + \sigma_\varepsilon I)^{-1} \mathbf{y}$$

Full Model

- Our model is a convex combination of f_i

$$y(x) = s_1(x) f_1(x) + \ldots + s_r(x) f_r(x) + \varepsilon_n$$

Switching functions

$$s_i(x) \in \Delta^r$$

Functional regimes

Model part 1: Functional Regimes

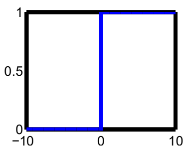
- GP prior for each functional regime
  - Use flexible stationary kernels

$$f_i \sim GP(0, K_i),\ \ i = 1, \ldots, r$$

Model part 2: Change Surfaces

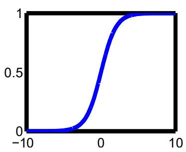
- Changepoint

$$s_i = I(t < T_i)$$



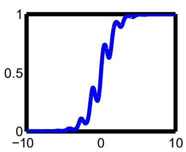
- Non-discrete changepoint

$$s_i = \text{softmax}(t - T_i)$$



- Change surface

$$s_i = \text{softmax}(w_i(t))$$
$$s_i = \sigma(w_i(t))$$



Model part 2: Change Surfaces

- Random Kitchen Sink features for $w_i(x)$
  - Variable rate of change
  - Non-monotonic
  - Heterogeneous over input

$$w_i(x) = \sum_{j=1}^{q} a_j \cos(\omega_j^T x + b_j)$$

Full Model

- Gaussian process change surface model

$$y(x) = \sum_{i=1}^{r} \sigma(w_i(x))\, f_i(x) + \varepsilon_n$$

$$f_i(x) \sim GP(0, K_i)$$

- Can depict this as a single Gaussian process with covariance function

$$k_{all}(x, x') = \sum_{i=1}^{r} \sigma(w_i(x)) k_i(x, x') \sigma(w_i(x'))$$

Scalable Inference

- Log likelihood naively O(n^3)

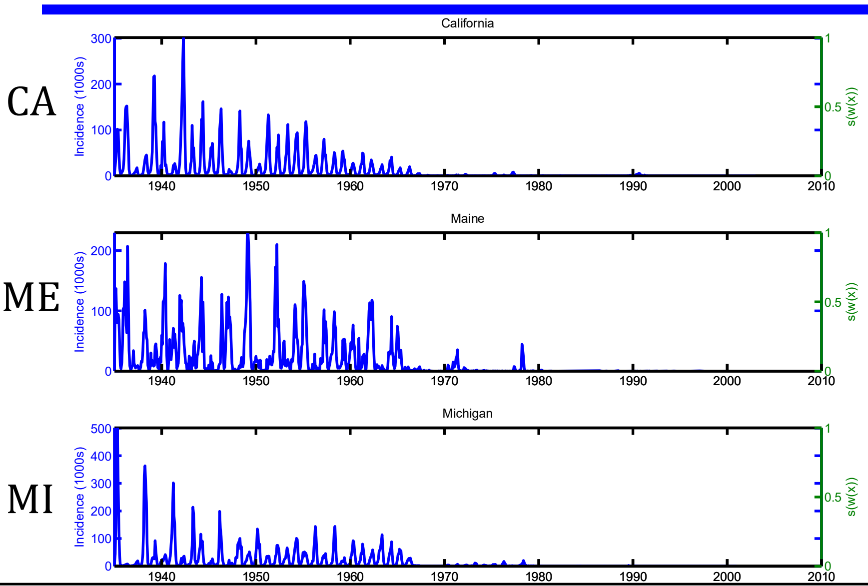
$$\log p(\mathbf{y} \mid \theta) \propto -\log \mid K + \sigma_\varepsilon I \mid - \mathbf{y}^T (K + \sigma_\varepsilon I)^{-1} \mathbf{y}$$

- We develop scalable Kronecker inference using the Weyl bound, O($Dn^{D+1/D}$)

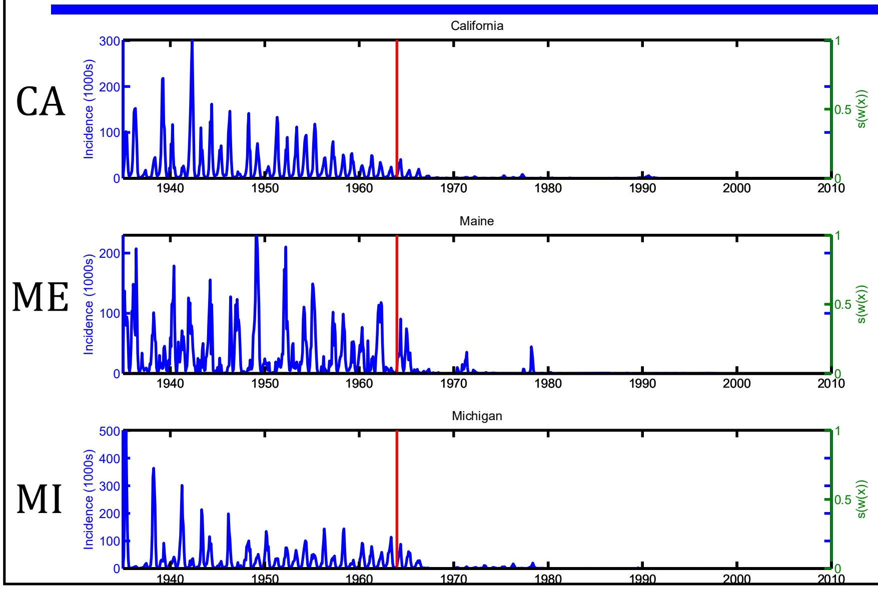
Measles in the United States

- Data
  - Monthly incidence rates 1935 – 2003
  - Continental United States and D.C.
  - $x \in \Re^3$, 2D space and 1D time
  - Measles vaccine introduced in 1963

Measles in 3 states



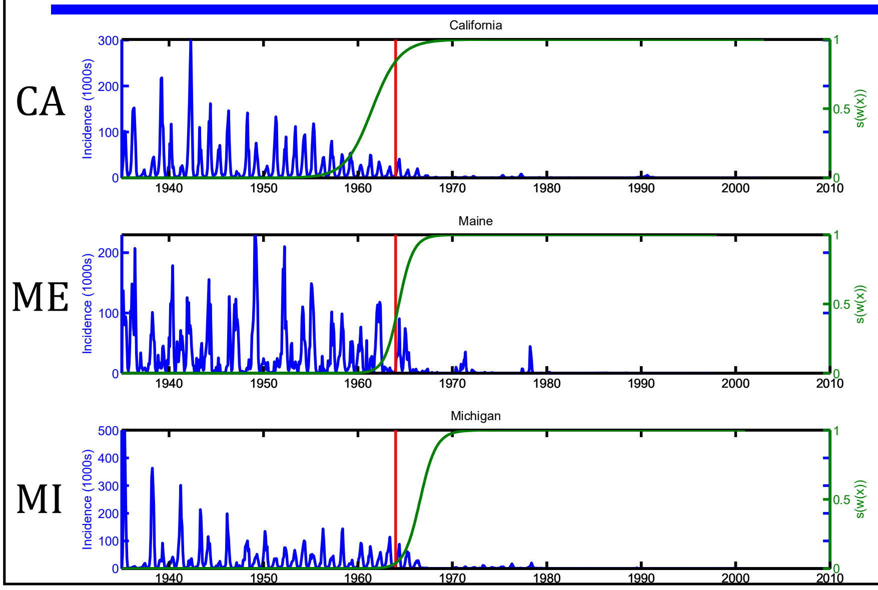
Measles in 3 states



Measles in 3 states

- GP change surface
  - 2 functional regimes
  - $w_i(x)$ as RKS with 5 features
- **Not a causal model!**

Measles in 3 states

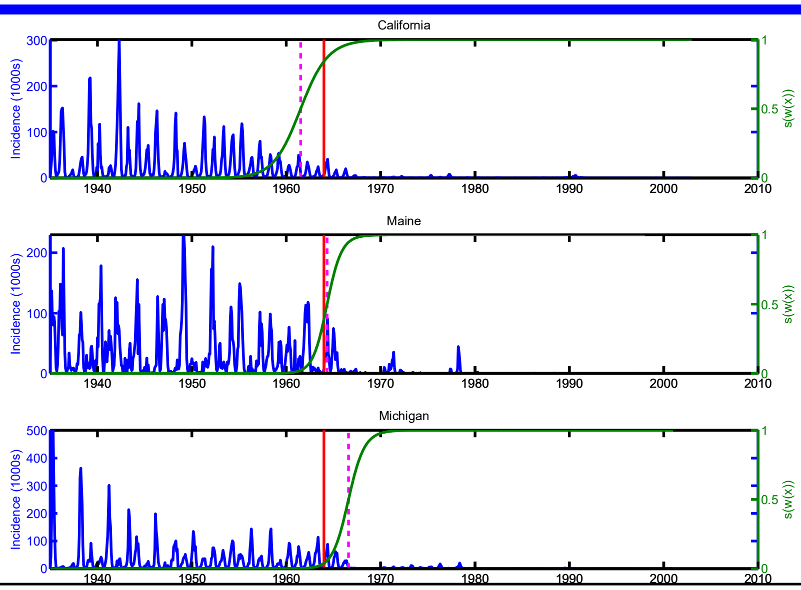


Measles in 3 states

CA

ME

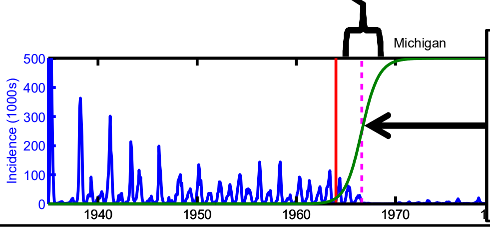
MI



Measles in 3 states

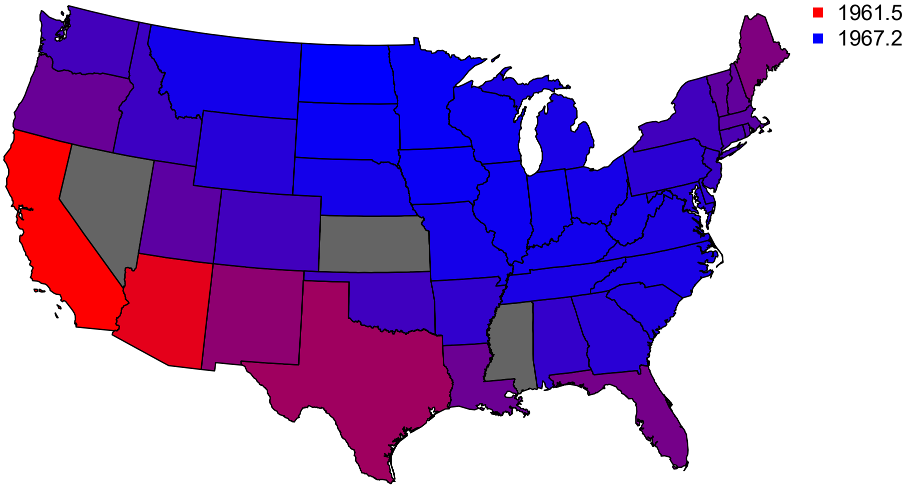
“Change slope” from
σ(w(x)) = 0.25 → 0.75

MI



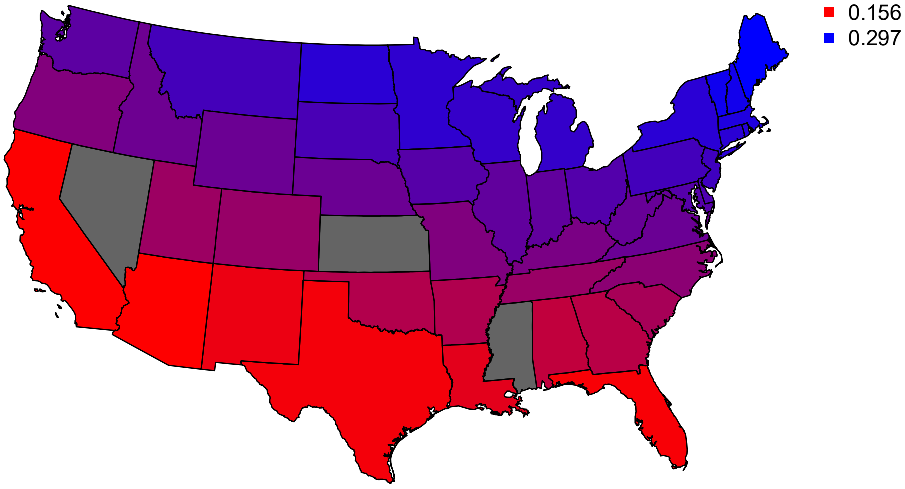
“Change date” per state
σ(w(x)) = 0.5

Change date for measles in U.S.



For each state, date where σ(w(x)) = 0.5

Change slope for measles in U.S.



For each state, slope of σ(w(x)) = 0.75 → 0.25

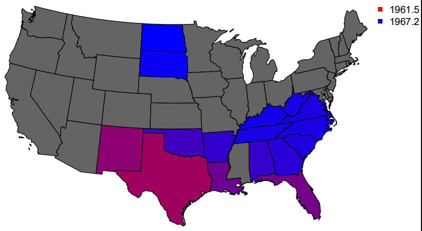
Regression Analysis

- Explore factors that affect the change date
  - Birth and death rates
  - Population numbers per age segment
  - Income information
  - Government hospital and health workers
  - Slope of change surface
  - Average temperature

	Dependent variable: Change date
Average date rate 0-4	−228.924
	(687.066)
Average date rate 5-9	8,928.639
	(6,153.325)
Average birth rate	−0.210
	[illegible]
Gini of family income	32.317*
	(12.071)
[illegible]	[illegible]
	[illegible]
Slope of change surface	37.913**
	(8.976)
[illegible]	[illegible]
	(165.077)
Population 0-4	−0.00002
	(0.00003)
Population 5-9	0.00002
	(0.00003)
Average temperature (°F)	0.025
	(0.041)
Constant	1,946.783**
	(7.614)
Observations	46
R^2	0.618
Adjusted R^2	0.446
Note:	*p<0.05; **p<0.01

Regression Analysis

- Gini of family income
  - Economically depressed communities
  - Rural regions



- Slope of change surface
  - Fewer cases nationwide enable more effective immunization later

Conclusions

- Introduced model for “change surfaces” in real world data
- Developed scalable inference for additive, non-stationary Gaussian processes
- Identified heterogeneity in first years of the measles vaccine
- Used the results of the change surface model for policy relevant conclusions

Acknowledgements

- Committee
  - Daniel Neill, Alex Smola, Wilbert van Panhuis
- Chair
  - Dave Choi
- Collaborators*
  - Andrew Wilson
  - Seth Flaxman
  - Hannes Nickisch

*Subset of paper accepted to AISTATS 2016

Questions?

Fin.

28

Backup slides

Conclusions

- Introduced model for “change surfaces” in real world data
- Developed scalable inference for additive, non-stationary Gaussian processes
- Identified heterogeneity in first years of the measles vaccine
- Used the results of the change surface model for policy relevant conclusions

Spectral Mixture Kernels

$$\sum_{q=1}^{Q} \omega_q cos(2\pi(\tilde{x}-\tilde{x}')^\top m_q) \prod_{p=1}^{P} \exp(-2\pi^2(\tilde{x}_p-\tilde{x}'_p)^2 v_q^{(p)})$$

Algorithm 2 Initialize spectral mixture kernels

for $k_i : i = 1 : r$ **do**
  for $d = 1 : D$ **do**
    Compute $x^{(d)} \in \{x : \sigma(w_i(x)) > 0.5\}$
    Sample $s \sim |FFT(sort(y(x^{(d)})))|^2$
    Fit Q component 1D GMM to s
    Initialize $\omega_q = std(y) * \phi_q$; $m_q = \mu_q$; $v_q = \sigma_q$
  end for
end for

Inference

- Compute log marginal likelihood

$$\log p(\mathbf{y}|\theta) \propto -\log|K + \sigma_\epsilon I| - \mathbf{y}^\top (K + \sigma_\epsilon I)^{-1}\mathbf{y}$$

- General Kronecker methods for scalability
  - Assume: $x \in X = X^{(1)} \times \ldots \times X^{(D)}$
  - Assume: multiplicative kernel across D
  - Then we can decompose kernel matrix,

$$K = K_1 \otimes \cdots \otimes K_D$$

Inference

- For additive kernels

$$K_i = K_1 \otimes \ldots \otimes K_D$$

$$K = \sum_{i=1}^{r} K_i = \sum_{i=1}^{r} K_{i,1} \otimes \ldots \otimes K_{i,D}$$

- K^{-1} can be computed efficiently using LCG*
- But how can we compute the log|K| ?

*See Flaxman et al. (2015)

Inference

(Weyl, 1912) which states that for $n \times n$ Hermitian matrices, $M = A + B$, with sorted eigenvalues $\mu_1, \ldots, \mu_n$, $\alpha_1, \ldots, \alpha_n$, and $\beta_1, \ldots, \beta_n$, respectively,

$$\mu_{i+j-1} \leq \alpha_i + \beta_j$$

$$\boxed{\begin{aligned} &\log(|A+B|) \\ &= \log(|M|) \\ &= \sum_{i=1}^{n} log(\mu_i) \\ &\leq \sum_{i+j-1=1}^{n} \log(\alpha_i + \beta_j) \end{aligned}}$$

Inference

- Choosing indices i, j $\sum_{i+j-1=1}^{n} \log(\alpha_i + \beta_j)$

Method	Complexity
Minimization for best pair	$O(n^2)$
"Middle" heuristic i=j OR i=j+1	O(n)
Greedy search of *s* pairs below and above previous pair	O(2sn)

Inference

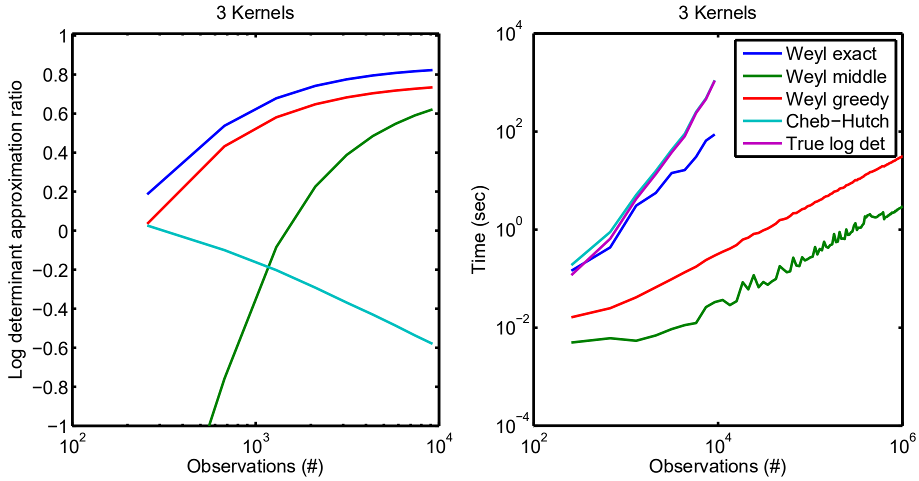
- Scaling functions, σ(w(x))

$$K = S_1 K_1 S_1' + \cdots + S_r K_r S_r' \tag{22}$$

where $S_i = \mathrm{diag}(\sigma(w_i(x)))$ and $S_i' = \mathrm{diag}(\sigma(w_i(x')))$. Employing the bound on eigenvalues of matrix products (Bhatia, 2013),

$$\mathrm{sort}(\mathrm{eig}(A * B)) \leq \mathrm{sort}(\mathrm{eig}(A)) * \mathrm{sort}(\mathrm{eig}(B)) \tag{23}$$

Inference



Inference – so what?!

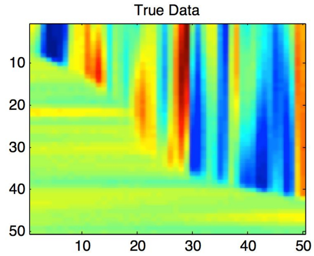
- Linear complexity for additive kernels
  - $O(Dn^{D+1/D})$
- Scalable inference for non-separable kernels in space and time
- Scalable inference for non-stationary kernels

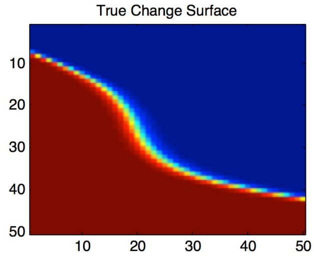
Numerical Experiments

- 2500 points of synthetic data
- 2 functional regimes defined by squared exponential kernels
- Change surface define by $\sigma(w_{poly}(x))$

$$w_{poly}(x) = \sum_{i=0}^{3} \beta_i^T x^i,\ \beta_i \sim \mathcal{N}(0, 3I_D)$$

Results - Numerical





	Dependent variable: Change slope
Average date rate 0-4	2.974
	(10.959)
Average date rate 5-9	−142.811
	(98.068)
Average birth rate	0.001
	(0.003)
Gini of family income	−0.531**
	(0.191)
Per capita income	−0.00000
	(0.00000)
Change date of change surface	0.010**
	(0.002)
Gov't health and hospitals employees per population	−5.007
	(2.644)
Population 0-4	0.00000
	(0.00000)
Population 5-9	−0.00000*
	(0.00000)
Average temperature (°F)	−0.002**
	(0.001)
Constant	−18.456**
	(4.483)
Observations	46
R^2	0.904
Adjusted R^2	0.860
Residual Std. Error	0.014 (df = 31)
F Statistic	20.784** (df = 14; 31)
Note:	*p<0.05; **p<0.01

Demographic Analysis

Coefficients:	Estimate	Std. Error	t value	Pr(> $\|t\|$)	
(Intercept)	1.944e+03	8.973e+00	216.688	<2e-16	***
Average date rate all	-2.572e+03	3.722e+03	-0.691	0.4948	
Average date rate 0-4	-2.967e+02	7.076e+02	-0.419	0.6780	
Average date rate 5-9	1.035e+04	6.809e+03	1.520	0.1389	
Average date rate 10+	2.052e+03	3.027e+03	0.678	0.5031	
Average birth rate	-2.016e-01	1.906e-01	-1.058	0.2986	
Gini of family income	3.834e+01	1.688e+01	2.271	0.0305	*
Per capita income	[illegible]	[illegible]	[illegible]	[illegible]	
Slope of change surface	3.211e+01	1.443e+01	2.225	0.0338	*
Gov't health and hospitals employees	3.320e+02	1.674e+02	1.984	0.0565	.
Population	8.026e-08	7.780e-06	0.010	0.9918	
Population 0-4	-1.931e-05	2.912e-05	-0.663	0.5124	
Population 5-9	2.677e-05	3.079e-05	0.870	0.3915	
Population 10+	-1.005e-06	8.503e-06	-0.118	0.9067	
Average temperature (°F)	-1.165e-03	6.519e-02	-0.018	0.9859	
Median household income	9.043e-05	1.749e-04	0.517	0.6089	