Bivariate Penalized Splines for Geo-Spatial Models

GuanNan Wang
Department of Mathematics
College of William & Mary

April 20, 2016

Motivation

Bivariate Splines

Triangulations

Bivatiate Penalized Spline Estimators

Partially Linear Bivariate Spline

IMAGE OF LENA SÖDERBERG

Damaged image with 50% of missing data observations

IMAGE OF LENA SÖDERBERG

Image recovered using thin-plate splines

IMAGE OF LENA SÖDERBERG

Face of Lena Söderberg with 8401 pixels.

Image of Lena Söderberg

Lena Söderberg from November 1972 issue of Playboy

Spatial Data and Modeling

- Spatial is relating to the position, area, shape, and size of things.
- Spatial describes how objects fit together in space.
- Data are facts and statistics collected together for inference and analysis.
- Spatial Data are data/information about the location and shape of, and relationships among, geographic features, usually stored as coordinates and topology.

SPATIAL MODEL

- A common goal in spatial modeling: predicting the value of a target variable Y over a two-dimensional domain.
- Let $\{\mathbf{X}_i = (X_{1i}, X_{2i})\}_{i=1}^n$ be a set of n points range over a bounded domain $\Omega \subseteq \mathbb{R}^2$ of an arbitrary shape.
- Let Y_i be the value of Y observed at point $\mathbf{X}_i$.
- Given n observations $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n = \{(X_{1i}, X_{2i}, Y_i)\}_{i=1}^n$, we assume

$$Y_i = m(\mathbf{X}_i) + \sigma(\mathbf{X}_i)\,\epsilon_i, \quad i = 1, \cdots, n$$

  - ϵ_i's are random errors and independent of $\mathbf{X}_i$;
  - m is an unknown smooth function.
- Goal: to estimate a function of m based on the n observations.

1-D SMOOTHING SPLINES

- The **smoothing spline** estimate of m is defined as a solution to the optimization problem:

$$\sum_{i=1}^{n}[Y_i - m(X_i)]^2 + \lambda \int [m''(t)]^2 dt$$

with λ as a fixed constant (roughness penalty parameter).

  - The 1st term ensures the closeness of the estimate to the data;
  - The 2nd term penalizes the curvature of the function;
  - Small $\lambda \Rightarrow$ an interpolating estimate;
  - Large $\lambda \Rightarrow m''(x) \to 0 \Rightarrow$ the least squares fit.

BIVARIATE SMOOTHING

Suppose we have two input variables X_1 and X_2.

- **Thin-plate spline smoother:** Wood (2003)
  - By penalizing the curvature of the spline surface, thin-plate spline is defined as a solution to the optimization problem

$$\sum_{i=1}^{n}(Y_i - m(X_{1i}, X_{2i}))^2 + \lambda \int \sum_{i+j=2} \binom{2}{i} (D_{x_1}^{i} D_{x_2}^{j} f)^2 dx_1 dx_2.$$

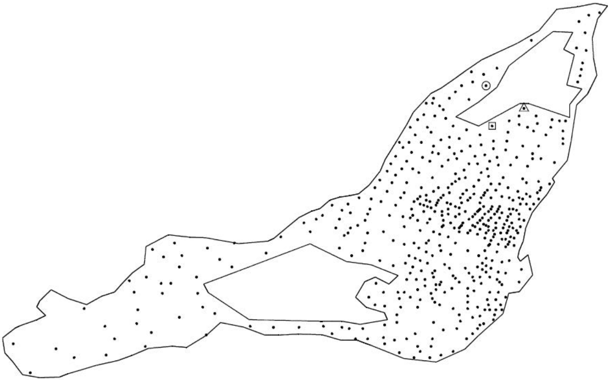
- **Tensor product spline smoother:**

$$m(x_1, x_2) = \sum_{j,k} \beta_{jk} B_j(x_1) B_k(x_2).$$

  - Useful when the data are observed on a regular grid in a rectangular domain;
  - Undesirable when data are located in domains with complex boundaries and holes.

Smoothing Over Difficult Regions

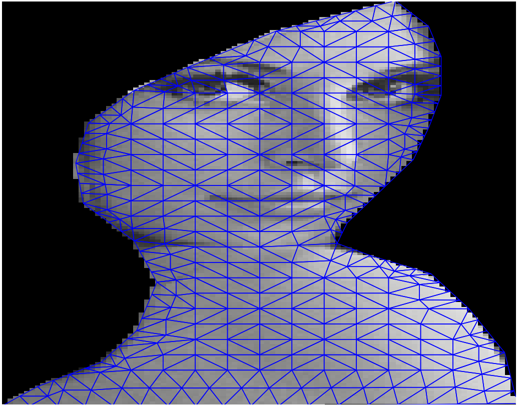
Ramsay (2002, JRSSB): estimate the per capita income for the Island of Montreal, Canada.



Island of Montreal with 493 data points defined by the centroids of census enumeration areas. Source: Ramsay (2002, JRSSB)

Bivariate Splines Over Triangulation

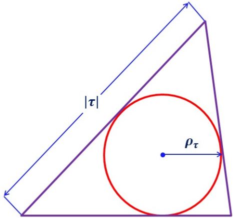
- We consider **bivariate splines on triangulations** to handle the irregular domains.



Triangulation of the image of Lena Söderberg

TRIANGLE: SIZE AND SHAPE

- Let τ be a triangle, i.e., a convex hull of three points not located in one line.
- Given any triangle τ,
  - Let $|\tau|$ be the length of its longest edge;
  - Let ρ_τ be the radius of the largest disk inscribed in τ;
  - Define the ratio $\beta_\tau = |\tau|/\rho_\tau$ the shape parameter of τ;
  - For an equilateral triangle, $\beta_\tau = 2\sqrt{3}$;
  - Any other triangle has a larger shape of parameter;
  - When β_τ is small, the triangles are relatively uniform (all angles of triangles in the triangulation τ are relatively the same).

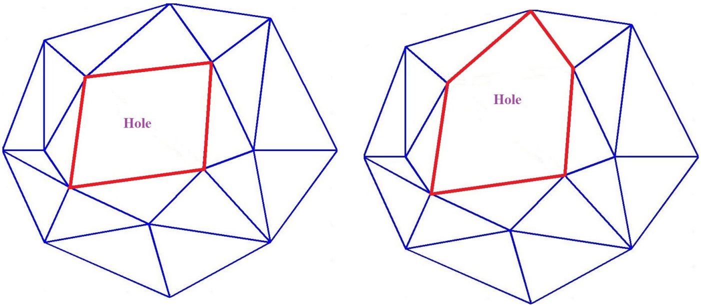


Triangulations

A collection $\triangle = \{\tau_1, ..., \tau_N\}$ of triangles is called a triangulation of $\Omega = \cup_{i=1}^{N} \tau_i$ if a pair of triangles in $\triangle$ intersect, then their intersection is either a common vertex or a common edge.

TRIANGULATIONS

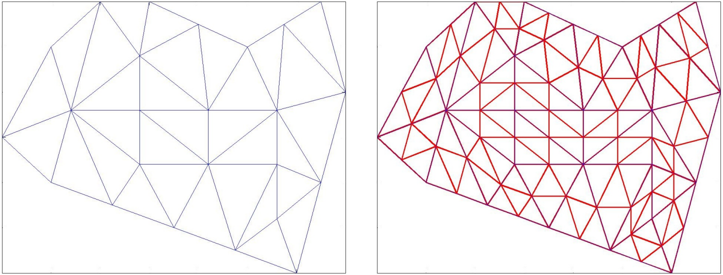
A collection $\triangle = \{\tau_1, ..., \tau_N\}$ of triangles is called a triangulation of $\Omega = \cup_{i=1}^{N} \tau_i$ if a pair of triangles in $\triangle$ intersect, then their intersection is either a common vertex or a common edge.



Two triangulation examples.

UNIFORM REFINEMENT OF A TRIANGULATION

Let Δ be a given triangulation. A uniform refinement of Δ can be obtained by splitting each triangle $\tau \in \Delta$ into four subtriangles by connecting the midpoints of the edges of τ.



A triangulation and its uniform refinement

Triangulations In Practice

- Maxmin-angle triangulation: we seek to maximize the smallest angle in a triangulation.
- There is no triangle that contains no data points.
- Find a polygon Ω containing all the design points of the data and triangulate Ω by hand or computer to have a triangulation $\triangle_0$.
- Uniformly refine $\triangle_0$ several times to have a desired triangulation.
- The Delaunay algorithm is a good way to triangulate the convex hull of an arbitrary dataset; see MATLAB program "delaunay.m".

Definition of Spline Functions

- Let $\tau = \langle (x_1, y_1), (x_2, y_2), (x_3, y_3) \rangle$. For any point $v = (x, y) \in R^2$, let b_1, b_2, b_3 be the solution of

$$\begin{aligned} x &= b_1 x_1 + b_2 x_2 + b_3 x_3, \\ y &= b_1 y_1 + b_2 y_2 + b_3 y_3, \\ 1 &= b_1 + b_2 + b_3, \end{aligned}$$

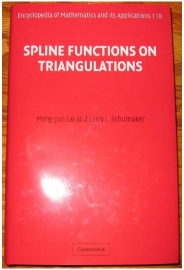
  where coefficients (b_1, b_2, b_3) are called the barycentric coordinates of point v with respect to the triangle τ.
- Fix a degree $d > 0$. For $i + j + k = d$, let

$$B^d_{ijk}(x, y) = \frac{d!}{i!j!k!} b_1^i b_2^j b_3^k \text{ (Bernstein-Bézier polynomials).}$$

- Let $s|_\tau = \sum_{i+j+k=d} c^\tau_{ijk} B^d_{ijk}(x, y)$, $\tau \in \triangle$.

Spline Functions on Triangulations

- Lai and Schumaker (2007): all basics about multivariate splines
  - Evaluation
  - Differentiation
  - Integration
  - Refinement schemes of a triangulation
- Lai and Schumaker (2007): advanced properties
  - Dimension of various spline spaces
  - Construction of various locally supported basis functions
  - Approximation properties of various spline spaces



Lai and Schumaker (2007, Cambridge Univ. Press)

BIVARIATE PENALIZED SPLINE ESTIMATOR

- Given $\lambda > 0$ and $\{\mathbf{X}_i, Y_i\}_{i=1}^n$, consider the minimization:

$$\min_s \sum_{i=1}^n \{Y_i - s(\mathbf{X}_i)\}^2 + \lambda \mathcal{E}_v(s), \tag{1}$$

where

$$\mathcal{E}_v(f) = \sum_{\tau \in \triangle} \int_\tau \sum_{i+j=2} \binom{2}{i} (D_{x_1}^i D_{x_2}^j f)^2 dx_1 dx_2$$

is the energy functional.

- Let $\widehat{m}_\lambda$ be the minimizer of (1) and we call it the bivariate penalized spline estimator over triangulation (BPSOT estimator) of m corresponding to λ.

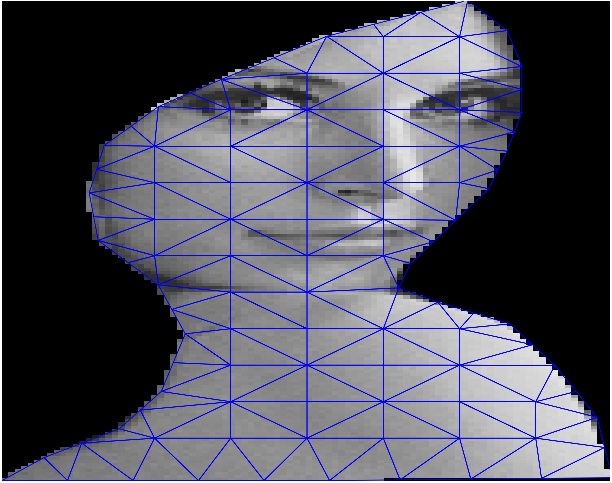
Penalty Parameter Selection

- Partition the original data randomly into K subsamples with: one subsample $\Rightarrow$ test set, $K-1$ subsamples $\Rightarrow$ training set.
- Define the K-fold cross-validation score as

$$CV_\lambda = \sum_{i=1}^{n} \left\{ Y_i - \hat{m}_\lambda^{-k[i]}(\mathbf{X}_i) \right\}^2$$

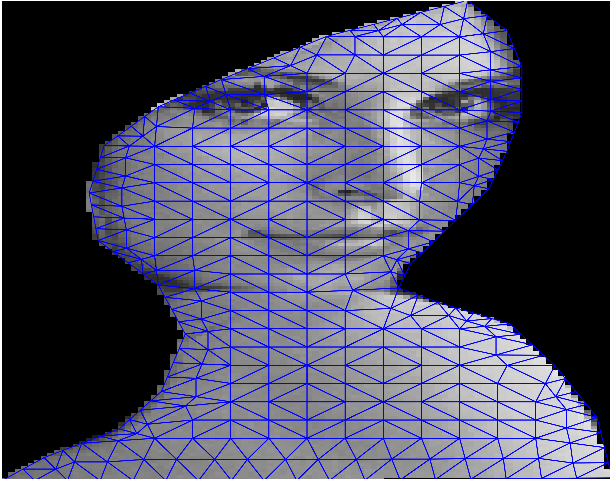
  - $k[i]$: the subsample containing the ith observation.
  - $\hat{m}_\lambda^{-k[i]}$: the estimate of the mean with the measurements of the $k[i]$th part of the data points removed.
- Select $\lambda = \arg\min CV_\lambda$.

IMAGE OF LENA SÖDERBERG: TRIANGULATIONS



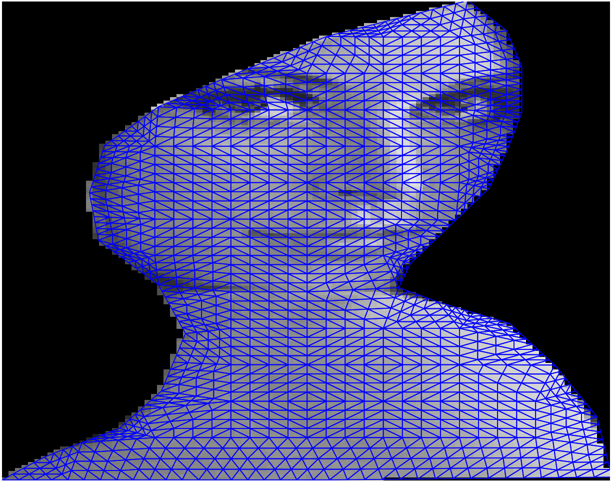
Triangulation $\triangle_0$

Image of Lena Söderberg: Triangulations



Triangulation $\triangle_1$

Image of Lena Söderberg: Triangulations

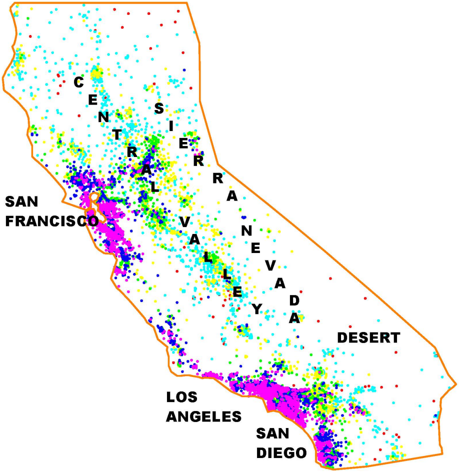


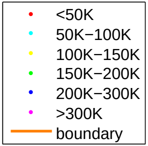
Triangulation $\triangle_2$

IMAGE OF LENA SÖDERBERG: TRIANGULATIONS

Recovered image using bivariate splines over triangulation Δ_1

Motivation: California House Value Data





- Prediction
- How different factors affect real estate property prices?

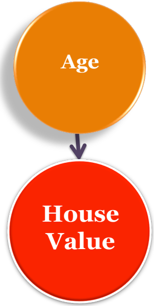
20,532 blocks defined by centroids of census enumeration areas (1990 Census).

CALIFORNIA HOUSE VALUE DATA

- **Data:** all the block groups in California from the 1990 Census
- **Target:** House value

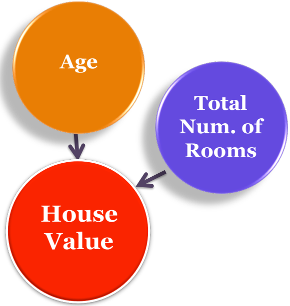
CALIFORNIA HOUSE VALUE DATA

- **Data:** all the block groups in California from the 1990 Census
- **Target:** House value



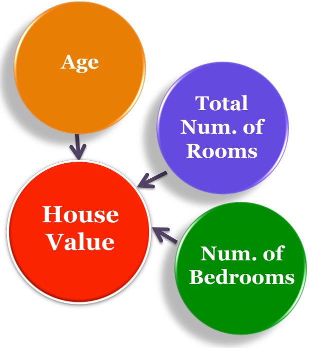
California House Value Data

- **Data:** all the block groups in California from the 1990 Census
- **Target:** House value



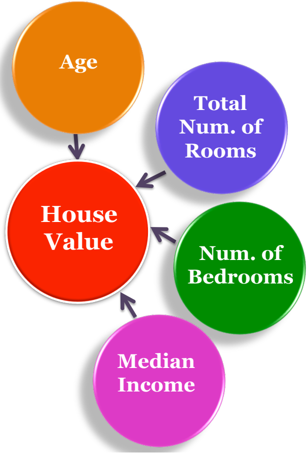
CALIFORNIA HOUSE VALUE DATA

- **Data:** all the block groups in California from the 1990 Census
- **Target:** House value



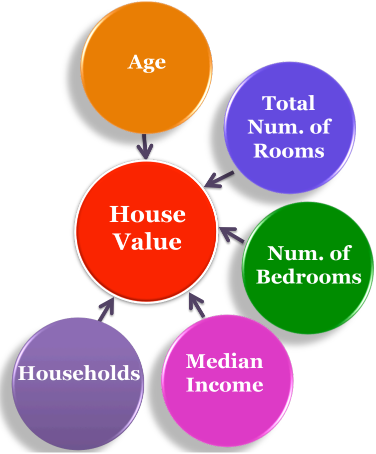
CALIFORNIA HOUSE VALUE DATA

- **Data:** all the block groups in California from the 1990 Census
- **Target:** House value



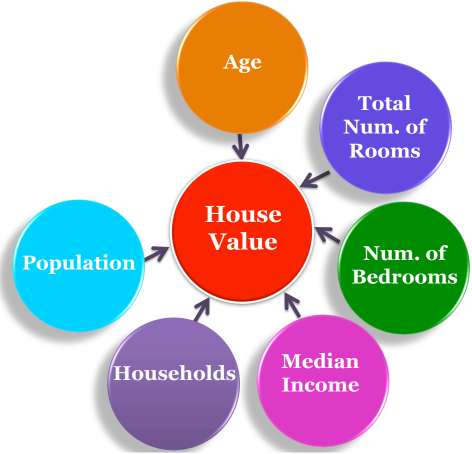
CALIFORNIA HOUSE VALUE DATA

- **Data:** all the block groups in California from the 1990 Census
- **Target:** House value



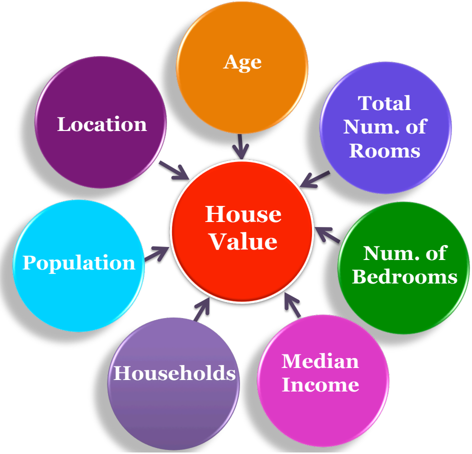
California House Value Data

- **Data:** all the block groups in California from the 1990 Census
- **Target:** House value



CALIFORNIA HOUSE VALUE DATA

- **Data:** all the block groups in California from the 1990 Census
- **Target:** House value



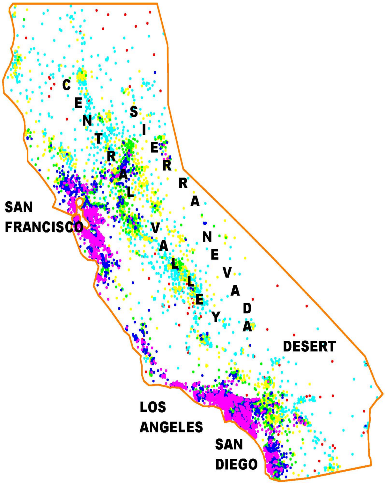
6-FACTOR GLM

- 6-factor GLM of house value as a linear combination of:
  - House Age (Age)
  - Total # of Rooms (TR)
  - # of Bedrooms (BR)
  - Median Income (Income)
  - Population (Pop)
  - Household (Hhd)

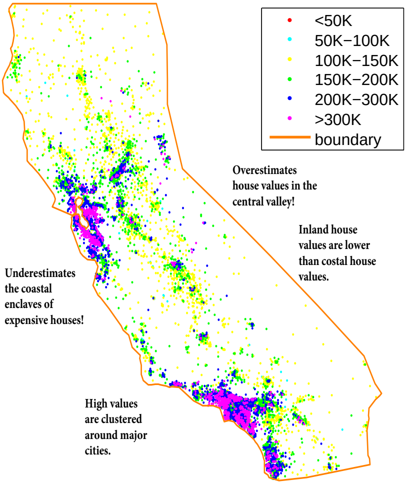
Model 1: 6-Factor GLM (Pace and Barry, 1997)

$$\begin{aligned}\log(\text{Value}) &= \beta_0 + \beta_1 \text{Income} + \beta_2 \log(\text{Age}) \\ &\quad + \beta_3 \log(\text{TR}/\text{Pop}) + \beta_4 \log(\text{BR}/\text{Pop}) \\ &\quad + \beta_5 \log(\text{Pop}/\text{Hhd}) + \beta_6 \log(\text{Hhd})\end{aligned}$$

6-FACTOR GLM ESTIMATES



A. California house value data



B. GLM fit with 6 factors

LOCATION, LOCATION, LOCATION!

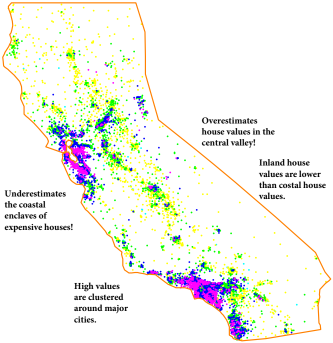
- "Location matters"!
- **We need to adjust for the location effect:**
  - House Age (Age)
  - Total # of Rooms (TR)
  - # of Bedrooms (BR)
  - Median Income (Income)
  - Population (Pop)
  - Household (Hhd)
  - Latitude
  - Longitude

Model 2: A Flexible Semiparametric Model

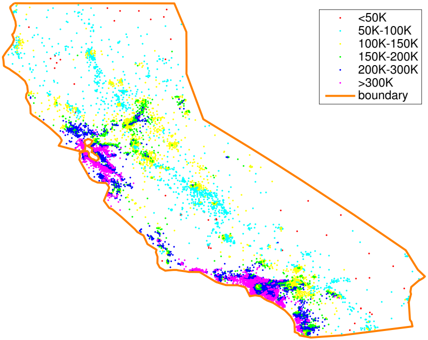
$$\begin{aligned} \log(\text{Value}) = {} & \beta_0 + \beta_1 \text{Income} + \beta_2 \log(\text{Age}) \\ & + \beta_3 \log(\text{TR}/\text{Pop}) + \beta_4 \log(\text{BR}/\text{Pop}) \\ & + \beta_5 \log(\text{Pop}/\text{Hhd}) + \beta_6 \log(\text{Hhd}) \\ & + g(\text{Latitude}, \text{Longitude}), \end{aligned}$$

where $g(\cdot, \cdot)$ is a smooth bivariate function to be estimated.

ESTIMATED HOUSE VALUES



(a) GLM fit with 6 factors



(b) Bivariate penalized splines

Prediction errors of the logarithm of house values.

LINEAR	KRIG	TPS	SOAP	BPST
0.146	0.083	0.081	0.079	0.052


thank you
danke
рахмат
謝謝
ngiyabonga
teşekkür ederim
gracias
merci
obrigado
dziękuję
bedankt
спасибо
Баярлалаа
sukriya
kop khun krap
go raibh maith agat
mochchakkeram
arigatô
dakujem
grazie
terima kasih
감사합니다
ありがとう
мерси
хвала
tapadh leat
dank je
takk
taiku
sagolun
hvala
mauruuru
faafetai lava
kiitos
dankie
dhanyavad
blagodaram
vinaka
mersi
barka
welalin
tack
spas
misaotra
matondo
paldies
grazzi
mahalo
asante
manana
obrigada
tenki
murakoze
chokrane
mamnun
trugarez
shukriya
dhanyavadagalu
diolch
tanemirt
rahmet
xiexie
ευχαριστώ
gracies
gratias ago
chnorakaloutioun
sulpay
djiere dieuf
najis tuke
kam sah hamnida
didi madloba
mèsi
dekuji
sobodi
enkosi
koszonom
köszönöm
nandri
nanni
gracie
bayarlalaa
akun
dankon
aciú
дякую
rahmat
merce

REFERENCES

- Furrer, R., Nychka, D., and Sainand, S. (2011). Package 'fields'. R package version 6.6.1. `http://cran.r-project.org/web/packages/fields/fields.pdf`.
- Lai, M. J. and Schumaker, L. L. (2007). *Spline Functions on Triangulations*. Cambridge University Press.
- Lai, M. J. and Wang, L. (2013). Bivariate penalized splines for regression. *Statistica Sinica* **23**, 1399-1417.
- Ramsay, T. (2002). Spline smoothing over difficult regions. *J R. Statist. Soc. B* **64**, 307-319.
- Wood, S. N. (2003). Thin plate regression splines. *J R. Statist. Soc. B*, **65**, 95-114.
- Wood, S. N., Bravington, M. V. and Hedley, S. L. (2008). Soap film smoothing.*J R. Statist. Soc. B* **70** 931-955.