Functional Change-point Detection for Multivariate Sparse Functional Data with Application on Urban Dynamics

Xinyue Chang, Xiongtao Dai and Zhengyuan Zhu   Iowa State University, Department of Statistics

Motivation

Detecting an urbanization process and estimating when the urbanization was happening is critical in the study of urbanization and will facilitate the development of urban growth models and the investigation of further environmental impacts.

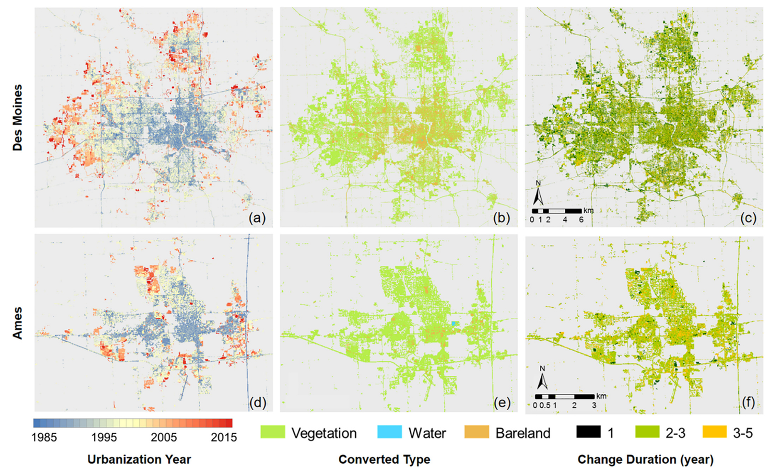

Figure 1: Annual dynamics of urban extents from 1985 to 2015 (a, d), associated conversion sources (b, e), and change duration (c, f) in Des Moines and Ames. [1]

Data Introduction

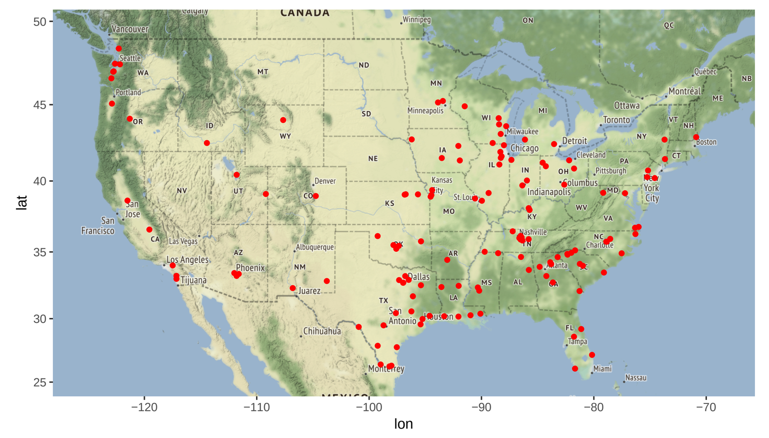

Figure 2: Some sampled locations

- 373 urbanized locations and 124 non-urbanized locations across the United States.
- For each location, surface reflectance data are available from 1985 to 2015 collected on 1–75 days. In each day, spectrum are recorded in 6 bands: B1 (Blue), B2 (Green), B3 (Red), B4 (NIR), B5 (MIR), and B7 (SWIR).

The urbanization change could be vegetation→urban, water→urban or bare land→urban, these changes are detected by NDVI = (NIR−Red)/(NIR+Red), MNDWI = (Green−MIR)/(Green+MIR), and SWIR(B7).

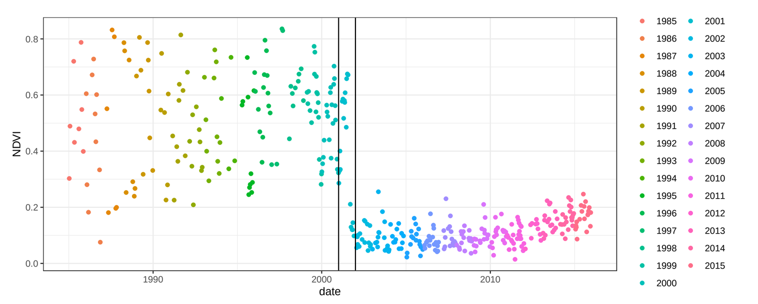

Figure 3: NDVI at 194th location

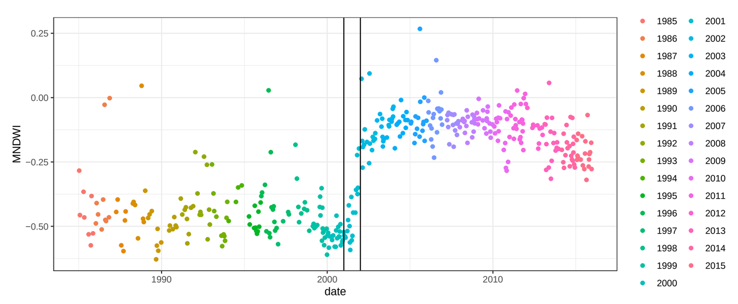

Figure 4: MNDWI at 194th location

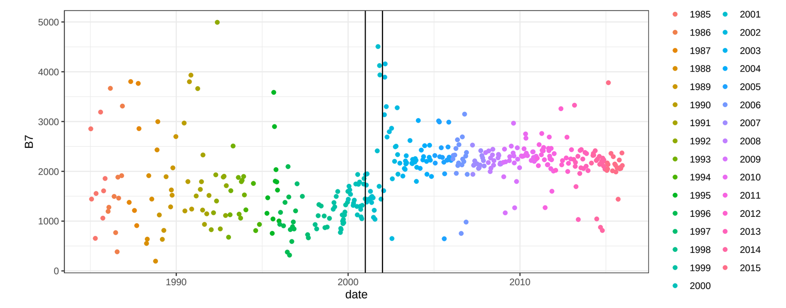

Figure 5: SWIR at 194th location

Multivariate Sparse Functional Data

For each location, model three indicators (or five bands) as the multivariate functional data with yearly replication. Let $i = 1, \cdots, N$ denote the year index, and t be the day of year,

$$\mathbf{Y}_i(t) = \mathbf{X}_i(t) + \boldsymbol{\epsilon}_i(t) = \boldsymbol{\mu}(t) + \sum_{k=1}^{K} \xi_{ik}\boldsymbol{\phi}_k(t) + \boldsymbol{\epsilon}_i(t)$$

where

- $\mathbf{Y}_i(t) = \{Y_{i1}(t), \cdots, Y_{ip}(t)\}$ corresponds to p observed time series in ith year.
- $\boldsymbol{\mu}(t) = \{\mu_1(t), \cdots, \mu_p(t)\}$ is the p-dimensional mean functions.
- $\boldsymbol{\phi}_k(t) = \{\phi_{k1}(t), \cdots, \phi_{kp}(t)\}$ is an orthonormal basis in Hilbert space.
- Principal component scores ξ_{ik} are uncorrelated with mean zero and variance λ_k.

We estimate ξ_{ik} and λ_k using RFPCA package [2], then work on time series of $\widehat{\xi}_{ik}$. K is determined by fraction of variation explained (FVE).

Functional Change-point Detection

Multivariate Weighted CUSUM Method (MWC):

$$T_N(x) = \frac{1}{N}\sum_{k=1}^{K}\widehat{\lambda}_k^{-1}\left(\sum_{1\le i\le Nx}\widehat{\xi}_{ik} - x\sum_{i=1}^{N}\widehat{\xi}_{ik}\right)^2, \quad 0 < x < 1$$

$$S_N = \frac{1}{N}\sum_{i=1}^{N} T_N\left(\frac{i}{N}\right) \xrightarrow{d} D_K^*$$

$$\widehat{\theta}_N = \underset{1\le l\le N}{\operatorname{argmax}}\ T_N\left(\frac{l}{N}\right)$$

Multivariate Max CUSUM Ensemble Method (MMCE):

$$T_N^k(x) = \frac{1}{N}\widehat{\lambda}_k^{-1}\left(\sum_{1\le i\le Nx}\widehat{\xi}_{ik} - x\sum_{i=1}^{N}\widehat{\xi}_{ik}\right)^2, \quad 0 < x < 1$$

$$S_N^k = \frac{1}{N}\sum_{i=1}^{N} T_N^k\left(\frac{i}{N}\right) \xrightarrow{d} D_1^*$$

$$\widehat{\theta}_N^k = \underset{1\le l\le N}{\operatorname{argmax}}\ T_N^k\left(\frac{l}{N}\right), \quad \widehat{\theta}_N = \widehat{\theta}_N^k \text{ where } k \text{ is the dominant component.}$$

Univariate Max Ensemble Method (UMWCE): (b denotes the band index)

$$T_N^b(x) = \frac{1}{N}\sum_{k=1}^{K_b}\widehat{\lambda}_k^{-1}\left(\sum_{1\le i\le Nx}\widehat{\xi}_{ik}^b - x\sum_{i=1}^{N}\widehat{\xi}_{ik}^b\right)^2, \quad 0 < x < 1$$

$$S_N^b = \frac{1}{N}\sum_{i=1}^{N} T_N^b\left(\frac{i}{N}\right) \xrightarrow{d} D_{K_b}^*$$

$$\widehat{\theta}_N^b = \underset{1\le l\le N}{\operatorname{argmax}}\ T_N^b\left(\frac{l}{N}\right), \quad \widehat{\theta}_N = \widehat{\theta}_N^b \text{ where } b \text{ is the dominant band index}$$

Univariate Sum Ensemble Method (USWCE):

$$T_N(x) = \sum_{b=1}^{p} T_N^b(x), \quad 0 < x < 1$$

$$S_N = \sum_{b=1}^{p} S_N^b, \quad \widehat{\theta}_N = \underset{1\le l\le N}{\operatorname{argmax}}\ T_N\left(\frac{l}{N}\right)$$

Method Illustration

To evaluate method performance, a repeated cross-validation is conducted as following,

Data → Preprocess data → Sparse MF-PCA: $\widehat{\xi}_{ik}$ and $\widehat{\lambda}_k$ → Repeat start → Split location sets into 10 folds → Estimate θ_N in one fold using FVE chosen in other 9 folds → RMSE $= \sqrt{\sum_{i=1}^{374}(\widehat{\theta}_N - \theta_N)^2}$ → Repeat end

To obtain the bootstrap distribution $D_K^*, K \ge 1$, we conducted the following procedure. For example, in MWC method,

Data → Estimate θ_N → $\mathbf{r}_i(t) = \mathbf{Y}_i(t) - \widehat{\boldsymbol{\mu}}_1(t)I(i \le \widehat{\theta}_N) - \widehat{\boldsymbol{\mu}}_2(t)I(i > \widehat{\theta}_N)$ → Repeat start → Resample 100 curves from $\mathbf{r}_i(t)$ → Calculate S_N → Stored as one sample from D_K^* → Repeat end

Change-point Testing

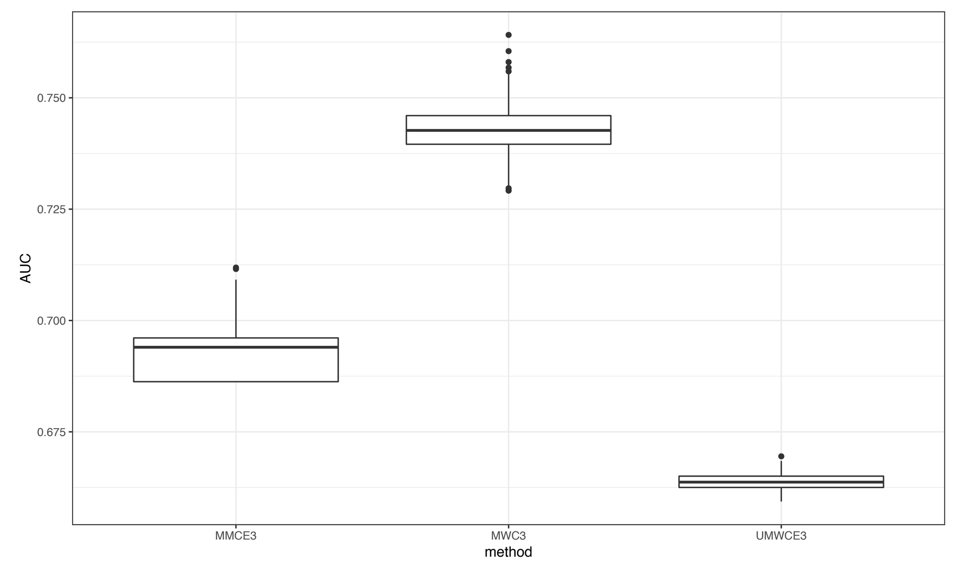

Figure 6: AUC (tuning α) boxplots for methods available. For the regression based method [1], AUC = 0.6377

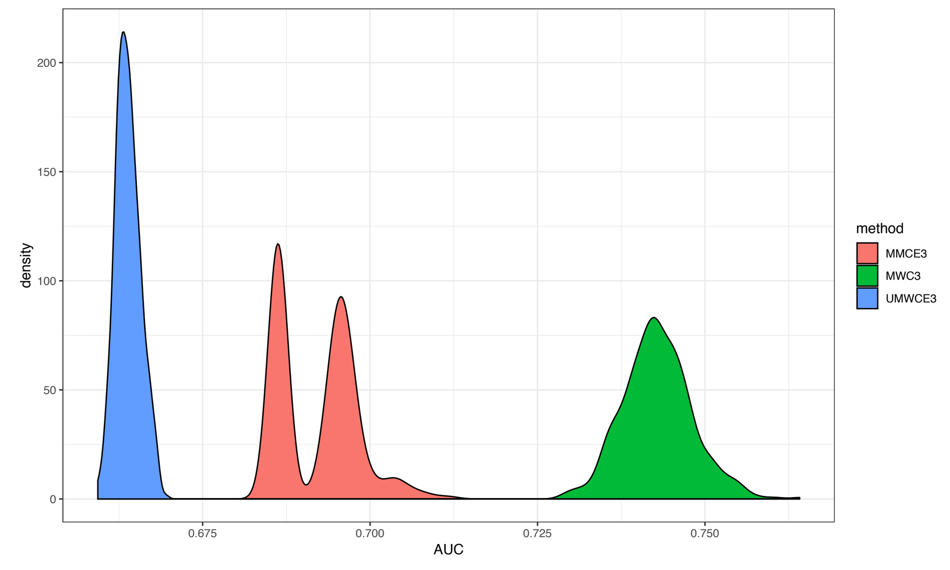

Figure 7: Density plots of AUC (tuning α) for methods available.

Change-point Estimation

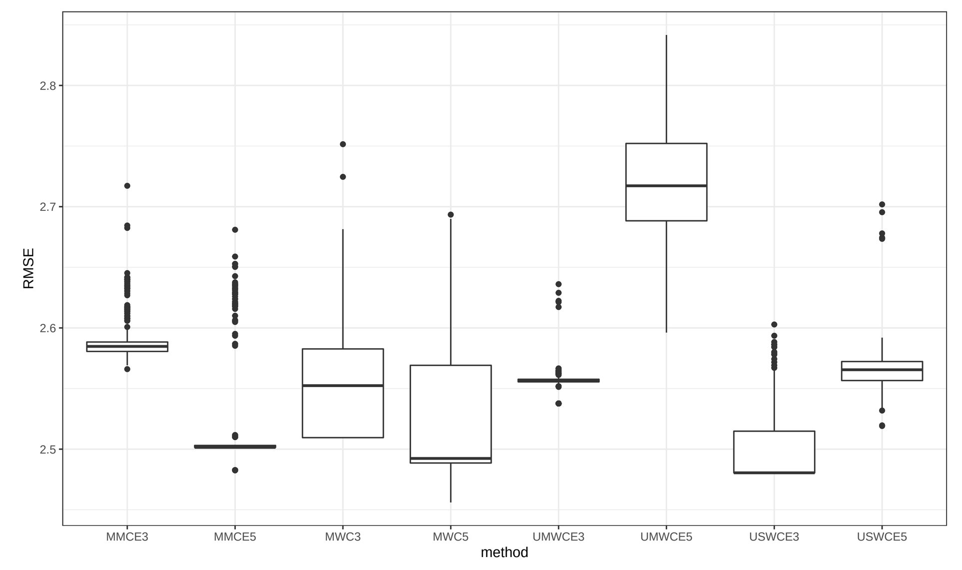

Figure 8: RMSE bxoplots for proposed different methods

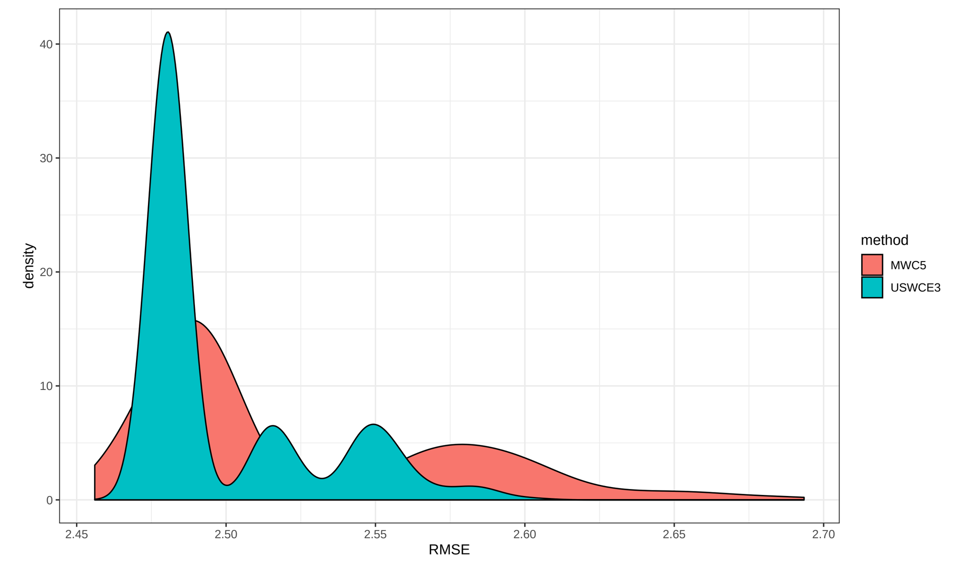

Figure 9: Comparison between two best methods.

Table 1: Average RMSE for different proposed methods. *Note: '3' denotes MFPCA using 3 indices to obtain $\widehat{\xi}_{ik}$ and '5' denotes MFPCA using 5 relevant bands instead.*

Method	MMCE3	MMCE5	MWC3	MWC5	UMWCE3	UMWCE5	USWCE3	USWCE5	Regression3	Regression5
Avg. RMSE	2.5895	2.5168	2.5516	2.5220	2.5548	2.7198	2.4990	2.5650	3.4556	3.3093

Conclusion

- The proposed methods all outperform a regression based benchmark method [1].
- Max ensemble methods are not sensitive to the FVE selection since it only utilizes the dominant component/index. So the ensemble is more stable compared to weighted methods.
- Except for USWCE3, multivariate method has great chance to be better than univariate method.

References

[1] Xuecao Li, Yuyu Zhou, Zhengyuan Zhu, Lu Liang, Bailang Yu, and Wenting Cao. Mapping annual urban dynamics (1985–2015) using time series of landsat data. *Remote Sensing of Environment*, 216:674–683, 2018.

[2] Xiongtao Dai, Zhenhua Lin, and Hans-Georg Müller. Modeling longitudinal data on riemannian manifolds. *arXiv preprint arXiv:1812.04774*, 2018.

[3] István Berkes, Robertas Gabrys, Lajos Horváth, and Piotr Kokoszka. Detecting changes in the mean of functional observations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):927–946, 2009.

Future Research

- Asymptotic properties of change-point testing statistics and estimation.
- Estimating multiple changes and gradual change.

Contact Information

Web https://chxyself25.github.io
Email xchang@iastate.edu
Phone +1 (218)7217292