Some spatial regression models for functional and compositional data

Tingting Huang[1], **Gilbert Saporta**[2],

[1] School of Statistics, Capital University of Economics and Business, Beijing, China

[2] CEDRIC Lab, CNAM, Paris, France

Conference in honour of Christine Thomas-Agnan

Outline

1. Introduction
2. Spatial regression with a functional predictor
   - 2.1 The functional linear model (SoFR)
   - 2.2 The spatial functional linear model (SSoFR)
   - 2.3 A robust version (RSSoFR)
   - 2.4 Application
3. Spatial regression models with a compositional predictor
   - 3.1 A spatial Durbin model with a compositional predictor and two competitors
   - 3.2 Application on real data
4. Conclusion and perspectives

1. Introduction

- Many regression models with spatially correlated data

$\mathbf{y}$ a vector of n observations of a response variable

$\mathbf{X}$ a matrix of *p* predictors

$\mathbf{W}$ an exogenous spatial (or neighbouring) weight matrix. Rows sum to 1 and diagonal elements are 0.

Morans'I statistic

$$I = \frac{\sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

- The general (non-identifiable) Manski model

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \mathbf{u} \quad \mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}$$

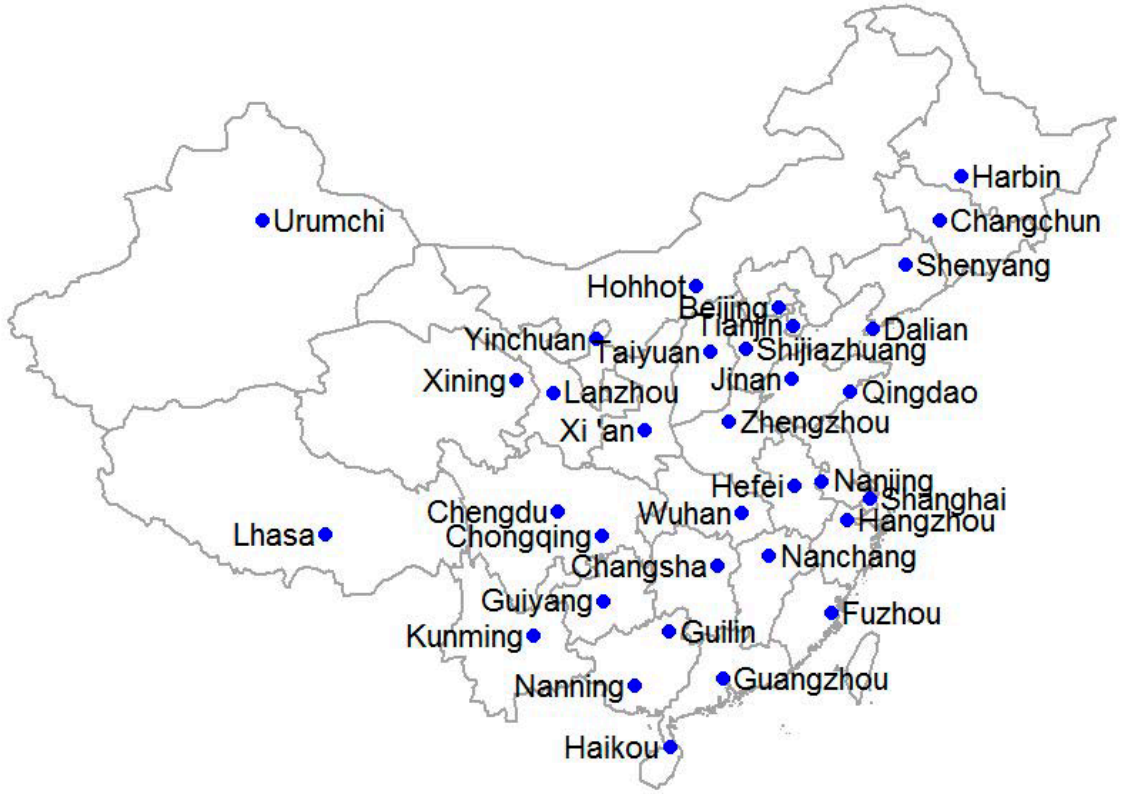
$$\boldsymbol{\varepsilon} \sim \mathrm{N}\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right)$$

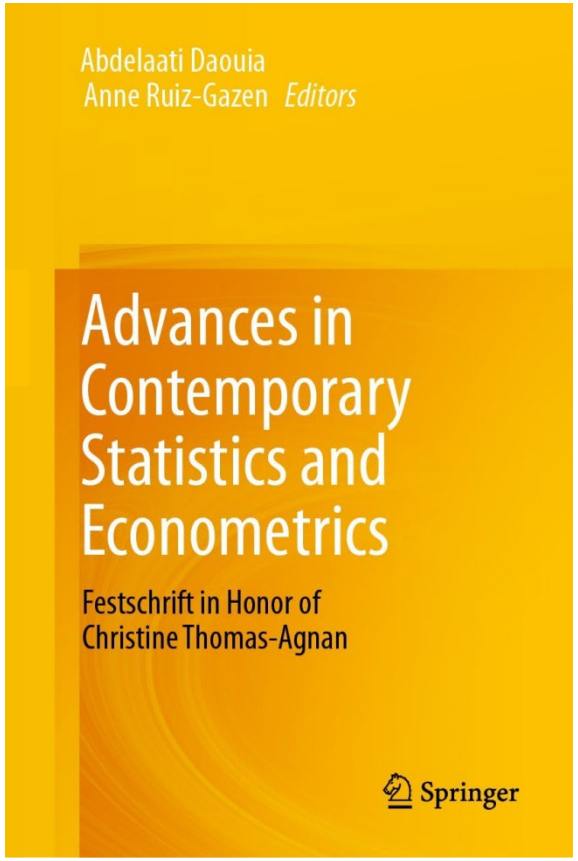
  - SAR model $\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
  - Spatial Durbin Model $\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$

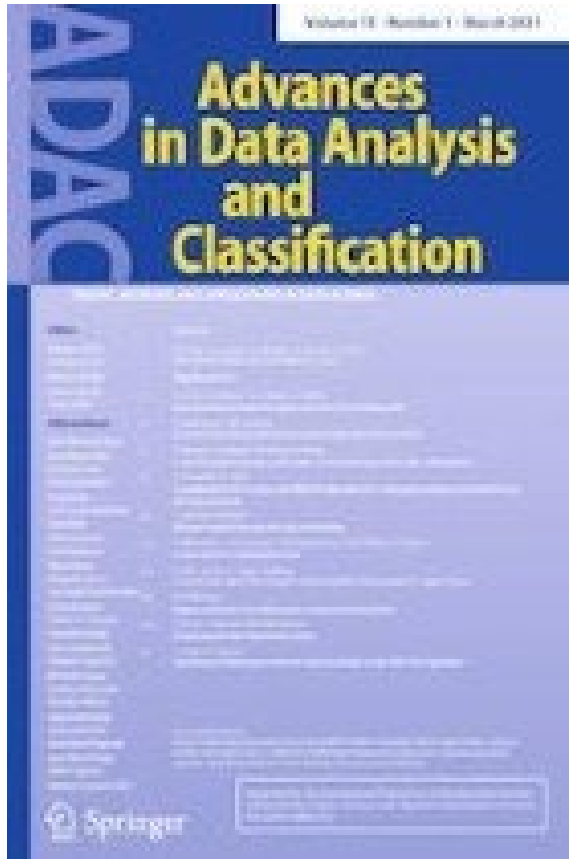
- ρ reflects the strength of spatial dependence

- Anselin, L. (1998). *Spatial econometrics: Methods and models*. Berlin: Springer.
- Elhorst, J. P. (2010). Applied spatial econometrics: raising the bar. *Spatial economic analysis*, *5*(1), 9-28.
- LeSage, J., & Pace, R. K. (2009). *Introduction to spatial econometrics*. Chapman and Hall/CRC.
- Loonis, V. *et al*. (eds) (2018). *Handbook of Spatial Analysis*, INSEE-Eurostat.

- This presentation:
  - Extensions of the SAR model for a functional predictor
  - An extension of the SDM for a compositional predictor
- Illustration with real data







Huang, T., Saporta, G., & Wang, H. (2021). A Spatial Durbin Model for Compositional Data. (pp. 471-488)

Huang, T., Saporta, G., Wang, H., & Wang, S. (2021). A robust spatial autoregressive scalar-on-function regression with t-distribution. *15*(1), 57-81.

2. Spatial regression with a functional predictor

$\mathbf{x}(t)$ a set of n time functions (or curves) observed on [0,T]

2.1 The Scalar on Function Regression model (SoFR)

$$\mathbf{y} = \int_0^T \beta(t)\mathbf{x}(t)dt + \boldsymbol{\varepsilon}$$

Ramsay, J.O. & Silverman, B.W. (1997); Cardot et al. (1999); Cai & Hall (2006); Hall & Horowitz (2007) etc.

R.A.Fisher « The Influence of Rainfall on the Yield of Wheat at Rothamsted »
Philosophical Transactions of the Royal Society, B: 213: 89-142 (1924)

Disregarding, then, both the arithmetical and the statistical difficulties, which a direct attack on the problem would encounter, we may recognise that whereas with q subdivisions of the year, the linear regression equations of the wheat crop upon the rainfall would be of the form

$$\bar{w} = c + a_1 r_1 + a_2 r_2 + \ldots + a_q r_q$$

where $r_1, r_2, \ldots, r_q$ are the quantities of rain in the several intervals of time, and $a_1, \ldots a_q$ are the regression coefficients, so if infinitely small subdivisions of time were taken, we should replace the linear regression function by a *regression integral* of the form

$$\bar{w} = c + \int_0^T a r \, dt, \qquad \ldots\ldots\ldots \quad \text{(III)}$$

where $r\,dt$ is the rain falling in the element of time dt; the integral is taken over the whole period concerned, and a is a *continuous* function of the time t, which it is our object to evaluate from the statistical data.

Minimizing $E\left(Y-\int_0^T \beta(t)X_t dt\right)^2$: Wiener-Hopf equations:

$$\operatorname{cov}(X_t, Y) = \int_0^T C(t,s)\beta(s)ds$$

An ill posed problem

- Picard's theorem states that $\beta(t)$ is unique if and only if: $\sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i^2} < \infty$

where the λ_i are the eigenvalues of the Karhunen-Loève expansion of X(t) and the c_i the covariances between *Y* and the functional principal components

$$X(t) = \sum_{i=1}^{\infty} f_i(t)\xi_i \qquad c_i = \operatorname{cov}(Y, \xi_i)$$

- Generally not true when n is finite and $\beta(t)$ is not.

Constrained solutions or **basis expansions**:

- Projection onto the *m* first components of:

  - Functional principal components regression (truncated KL expansion)

  - Functional PLS regression[1]

$$\max_{w}\left(\operatorname{cov}^2(Y,\int_0^T w(t)X_t dt)\right) \quad \|w\|^2 = 1$$

$$w(t) = \frac{\operatorname{cov}(X_t, Y)}{\sqrt{\int_0^T \operatorname{cov}^2(X_t, Y)dt}} \quad t = \int_0^T w(t)X_t dt$$

$$iteration\ under\ orthogonality$$

$$\hat{Y}_{PLS(q)} = c_1 t_1 + \ldots + c_q t_q = \int_0^T \hat{\beta}_{PLS(q)}(t)X_t \mathrm{dt}$$

$$R^2(Y;\hat{Y}_{PLS(q)}) \geq R^2(Y;\hat{Y}_{PCR(q)})$$

[1] Preda, C., Saporta, G.(2005) PLS regression on a stochastic process , *Computational Statistics & Data Analysis*, 48(1), pp. 149-158

2.2 The Scalar on Function Spatial Regression model (SSoFR)

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \int_0^T \beta(t)\mathbf{x}(t)dt + \boldsymbol{\varepsilon}$$

- SSoFR has merits of both SoFR and SAR. It reduces to the classical SoFR when ρ= 0
- A reformulation shows that the error terms are not independent:

$$\mathbf{y} = \left(\mathbf{I}_n - \rho \mathbf{W}\right)^{-1} \int_0^T \beta(t)\mathbf{x}(t)dt + \left(\mathbf{I}_n - \rho \mathbf{W}\right)^{-1} \boldsymbol{\varepsilon}$$

maximum likelihood should be used instead of OLS

- The spatial autocorrelation parameter ρ, the slope function β(t), and the variance of the error term σ^2 are estimated by Maximum Likelihood combined with a truncated functional PCR or PLS basis expansion:

$$\mathbf{y} \simeq \rho \mathbf{W}\mathbf{y} + \sum_{j=1}^{m} \mathbf{a}_j b_j + \boldsymbol{\varepsilon}$$

And its sample counterpart:

$$\mathbf{y} \simeq \rho \mathbf{W}\mathbf{y} + \sum_{j=1}^{m} \hat{\mathbf{a}}_j \hat{b}_j + \boldsymbol{\varepsilon}$$

$\mathbf{e} = \mathbf{y} - \boldsymbol{\rho}\mathbf{W}\mathbf{y} - \mathbf{A}\mathbf{b}$ normally distributed with variance $\sigma^2\mathbf{I}$

$$\ln L(\rho, \mathbf{b}, \sigma^2) = -\frac{n}{2}\ln\left(2\pi\sigma^2\right) + \ln\left|\mathbf{I} - \rho\mathbf{W}\right| - \frac{\mathbf{e}'\mathbf{e}}{2\sigma^2}$$

2.3 A Robust Scalar on Function Spatial Regression model (RSSoFR)

- Same formula:

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \int_0^T \beta(t)\mathbf{x}(t)dt + \boldsymbol{\varepsilon}$$

- But with different iid residuals $\varepsilon_i \sim \sigma T_\nu$
- Estimation by an EM algorithm using the following property:

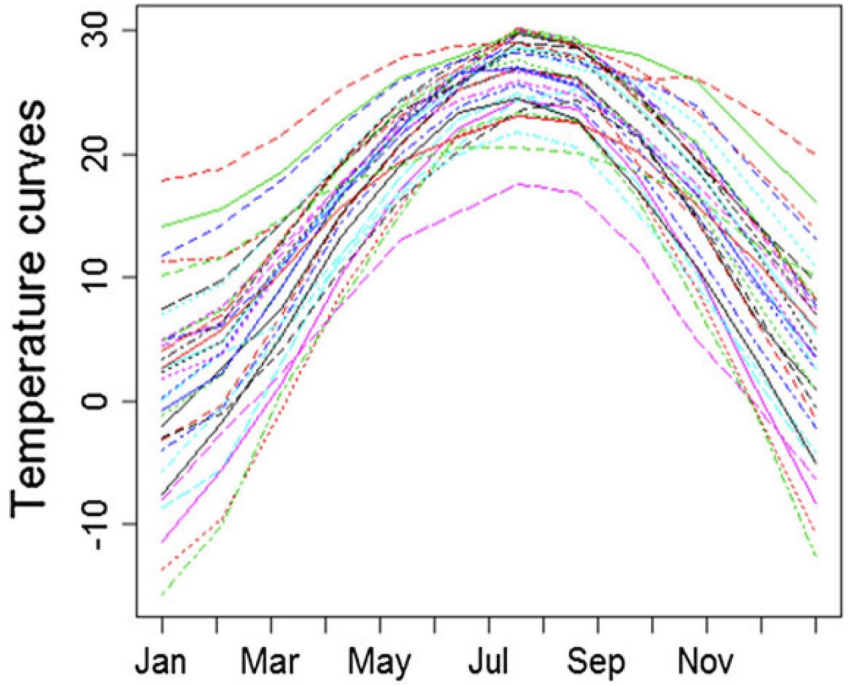
$$u_i \sim \gamma_{\nu/2} \quad \text{and} \quad \varepsilon_i \,|\, u_i \sim N\left(0, \frac{\sigma^2}{u_i}\right)$$

then $\dfrac{\varepsilon_i}{\sigma} \sim T_\nu$

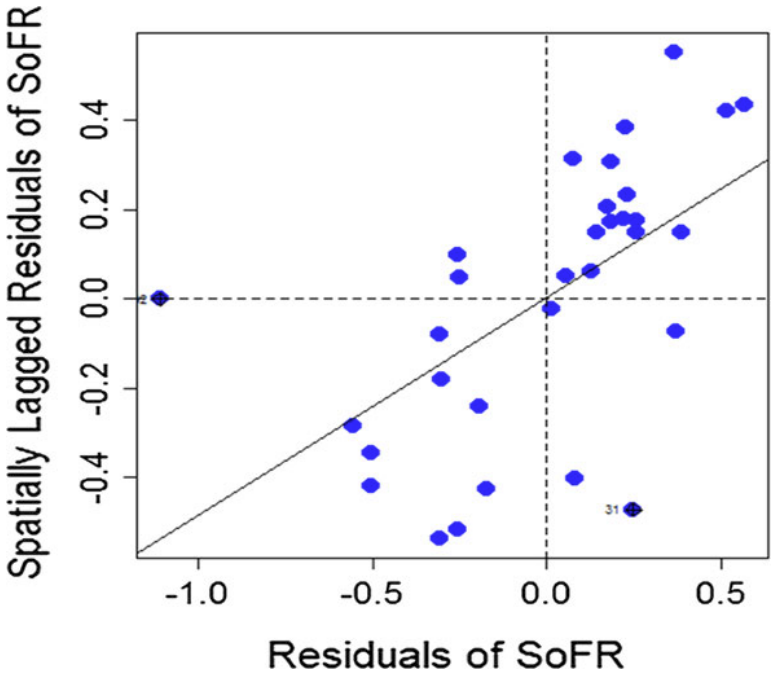
2.4 Application

- Data: monthly mean temperatures and total precipitation in 34 major cities in China 2005-2007. Aim: investigate the effect of temperature on precipitation over these 3 years. 2008 is the test set.

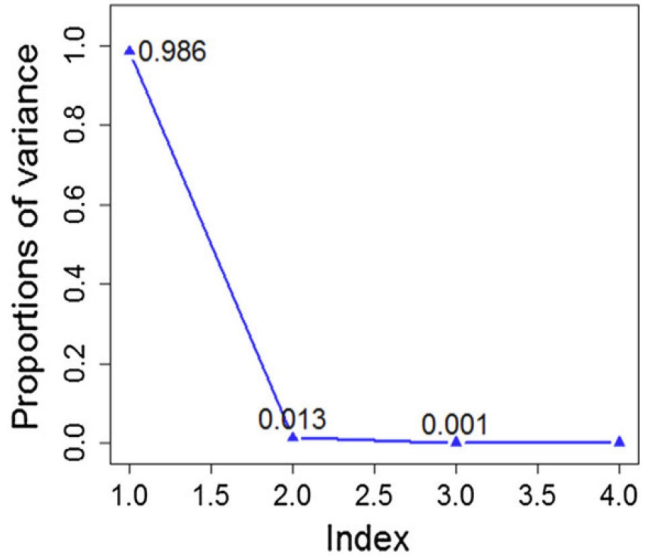
  - y_i logarithm of the mean annual total precipitation for the *i*th city.
  - Monthly temperature are smoothed by a kernel function



- Residuals shows a spatial pattern. The non-spatial SoFR does not fit to the data



- The spatial weight matrix $\mathbf{W}$ is formed by the reciprocal of the geodesic distance $d_{ii'}$ between centers of two cities i and i'
- A distance threshold is then applied since if city *i* is very far from city *i'* the spatial dependence between them will be very small. 626 km was found to be optimal according to Moran's statistic.
- Lhassa and Urumqi were removed.
- Only one component (PCR or PLS) was necessary



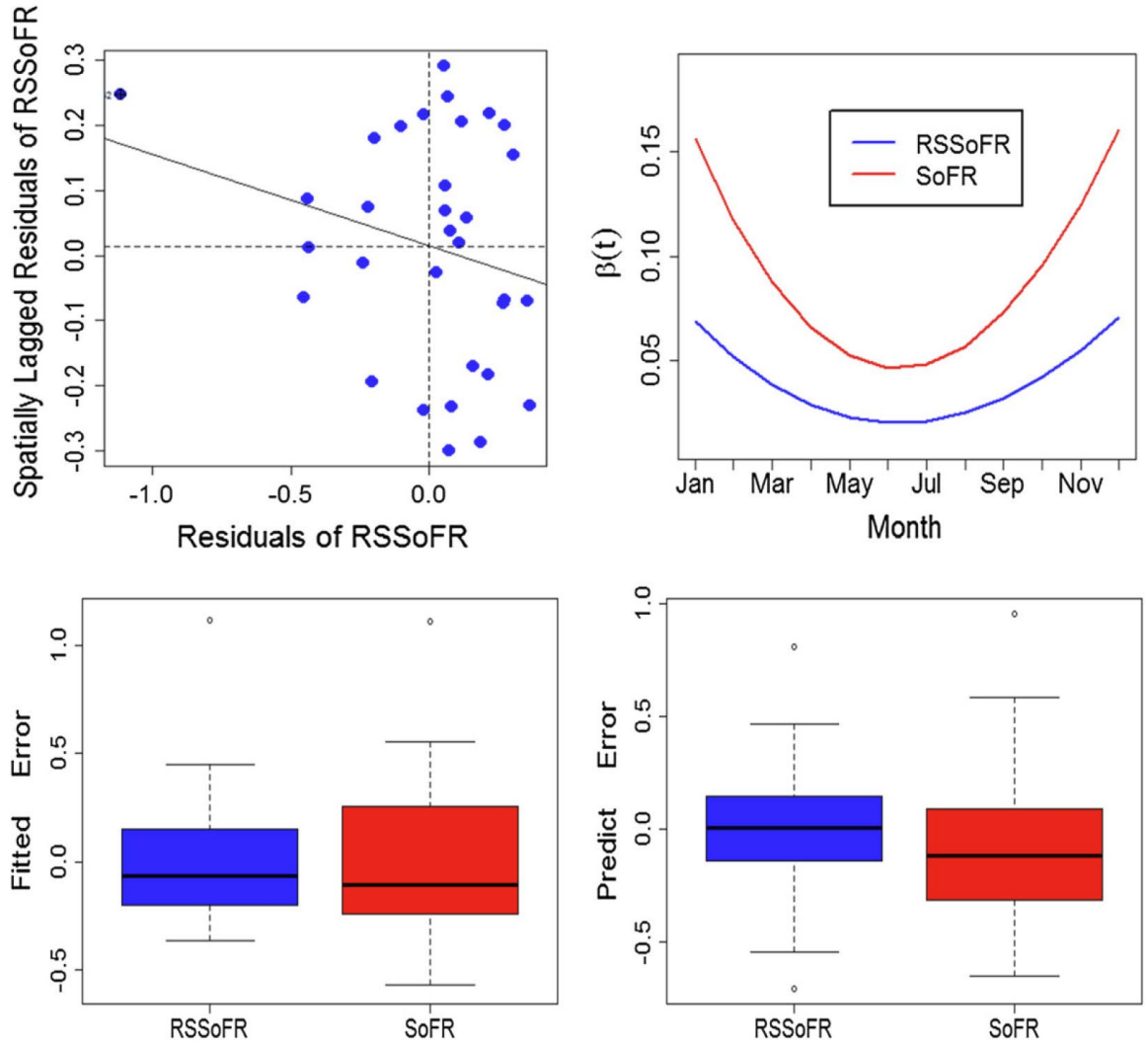


Fig. 5 The Moran Scatter plot of the residuals of the RSSoFR (top-left), the estimated $\beta(t)$ of the RSSoFR and the SoFR (top-right), the fitted error of the RSSoFR and the SoFR (bottom-left), and the predicted error of the RSSoFR and the SoFR (bottom-right), respectively

Precipitation is much more strongly influenced by the temperatures during spring and winter than in the other seasons

Comparison

MODEL	ρ	Moran's I statistic (residuals)	MSE (fitted error)	MSE (prediction error)	ν	Variance proportion
SoFR(PCA)	--	0.547	0.149	0.156	--	98.5%
SoFR(PLS)	--	0.487	0.128	0.114	--	82.9%
SSoFR (PCA)	0.80	0.107	0.092	0.131	--	98.5%
SSoFR (PLS)	0.76	0.053	0.089	0.078	--	82.9%
RSSoFR (PCA)	0.80	0.133	0.090	0.129	14	98.5%
RSSoFR (PLS)	0.675	-0.14	**0.088**	**0.076**	15	82.9%

3. Spatial regression models with a compositional predictor

3.1 A few notations

Composition: $\mathbf{x}^D = \left(x_1^D, \cdots, x_D^D\right)' \quad | \; x_j^D > 0, \; \sum_{j=1}^{D} x_j^D = 1$

Perturbation and powering:

$$\mathbf{x}^D \oplus \mathbf{y}^D = C\left(x_1^D y_1^D, x_2^D y_2^D, \cdots, x_D^D y_D^D\right) \qquad \alpha \odot \mathbf{x}^D = C\left((x_1^D)^\alpha, (x_2^D)^\alpha, \cdots, (x_D^D)^\alpha\right)$$

Inner Aitchison product:

$$\langle \mathbf{x}^D, \mathbf{y}^D \rangle_a = \sum_{j=1}^{D} \ln \frac{x_j^D}{g_m(\mathbf{x}^D)} \ln \frac{y_j^D}{g_m(\mathbf{y}^D)}$$

$$\langle \mathbf{X}^D, \boldsymbol{\theta}^D \rangle_a = \begin{pmatrix} \langle \mathbf{x}_1^D, \boldsymbol{\theta}^D \rangle_a \\ \vdots \\ \langle \mathbf{x}_n^D, \boldsymbol{\theta}^D \rangle_a \end{pmatrix}$$

$$\mathbf{W} \odot \mathbf{X}^D = \begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{pmatrix} \odot \begin{pmatrix} \mathbf{x}_1^D \\ \vdots \\ \mathbf{x}_n^D \end{pmatrix} = \begin{pmatrix} \bigoplus_{k=1}^{n} w_{1k} \odot \mathbf{x}_k^D \\ \vdots \\ \bigoplus_{k=1}^{n} w_{nk} \odot \mathbf{x}_k^D \end{pmatrix}$$

3.2 Models

- SAR Compositional Model (SARCD)[1]

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \langle \mathbf{X}^D, \boldsymbol{\beta}^D \rangle_a + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

- Spatial Error Model for Compositional Data (SEMCD)

$$\mathbf{y} = \langle \mathbf{X}^D, \boldsymbol{\beta}^D \rangle_a + \mathbf{u}, \quad \mathbf{u} = \rho \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

- Spatial Durbin Model for Compositional Data (SDMCD)

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \langle \mathbf{X}^D, \boldsymbol{\beta}^D \rangle_a + \langle \mathbf{W} \odot \mathbf{X}^D, \boldsymbol{\theta}^D \rangle_a + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

1 Huang, T., Wang, H., & Saporta, G. (2019). Spatial autoregressive model for compositional data. *Journal of Beijing University of Aeronautics and Astronautics*, 45(1), 93–98

- When $\rho = 0$ and $\boldsymbol{\theta}^D = \left(\frac{1}{D}, \frac{1}{D}, \cdots, \frac{1}{D}\right)'$ $\mathbf{y} = \langle \mathbf{X}^D, \boldsymbol{\beta}^D \rangle_a + \boldsymbol{\varepsilon}$

linear model with a compositional predictor

- When $\boldsymbol{\theta}^D = \left(\frac{1}{D}, \frac{1}{D}, \cdots, \frac{1}{D}\right)'$ $\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \langle \mathbf{X}^D, \boldsymbol{\beta}^D \rangle_a + \boldsymbol{\varepsilon}$

SAR Compositional Model (SARCD)

3.3 Estimation of the SDMCD

- First step : ilr [1] transformation for $\mathbf{x}^D$ and $\boldsymbol{\beta}^D$

$$\xi_j = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j^D}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k^D}}, \quad j = 1, 2, \cdots, D-1.$$

- Regression equation:

$$E\left(Y \mid \boldsymbol{\xi}\right) = \gamma_0 + \gamma_1^{(l)} \xi_1^{(l)} + \ldots + \gamma_{D-1}^{(l)} \xi_{D-1}^{(l)}$$

- [1] Hron, K., Filzmoser, P., & Thompson, K. (2012). Linear regression with compositional explanatory variables. *Journal of Applied Statistics*, *39*(5), 1115–1128

Then $\langle \mathbf{X}^D, \boldsymbol{\beta}^D \rangle_a = \boldsymbol{\Xi}\boldsymbol{\gamma}, \quad \langle \mathbf{W} \odot \mathbf{X}^D, \boldsymbol{\theta}^D \rangle_a = \mathbf{W}\boldsymbol{\Xi}\boldsymbol{\theta}$
and one may rewrite the compositional equation of SDMCD as a scalar equation:

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \boldsymbol{\Xi}\boldsymbol{\gamma} + \mathbf{W}\boldsymbol{\Xi}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

- Second step : Maximum Likelihood Estimation and back to the original coefficients
- Important note:
  - ξ_1 represents the relative information of x_1^D as regard to average of the rest of parts and $\gamma_1^{(l)}$ can be assigned to this part.
  - The remaining regression coefficients are not straightforward to interpret since the assigned regressor variables do not fully represent one particular part. Thus, the only way to interpret the role of each compositional part for explaining the response Y is to consider *D* different regression models by taking $l \in \{1, \ldots, D\}$, and to interpret the coefficient $\gamma_1^{(l)}$, representing part $x_1^{(l)}$

3.4 Application to real data

- SDMCD model is employed to investigate how three strata of industry* affect pollution of PM 2.5 (fine particulate matters with diameter smaller than 2.5 micrometer) in 34 major cities of China in 2016.
- Moran's I statistics = 0.63. Spatial effects are significant and there is a need to utilize spatial models.
- Weight matrix $\mathbf{W}$ is similar (but not identical, d_{ij}<459km) as in part 2

*proportions of the primary sector, the secondary sector and the tertiary sector in gross domestic product (GDP)

SARCD	R^2	Moran's I	ρ	Sector 1	Sector 2	Sector 3
	0.73	0.04	0.55	$\gamma_1^{(1)} = -2.8$	$\gamma_1^{(2)} = 12.99$	$\gamma_1^{(3)} = -10.2$
				$\hat{\beta}_1^D = 1.494 \times 10^{-6}$	$\hat{\beta}_2^D = 0.9999975$	$\hat{\beta}_3^D = 6 \times 10^{-9}$
		(0.33)	(0.0002)	(0.40)	(0.08)	(0.09)

SEMCD	R^2	Moran's I	ρ	Sector 1	Sector 2	Sector 3
	0.74	0.07	0.58	$\gamma_1^{(1)} = 1.32$	$\gamma_1^{(2)} = 6.46$	$\gamma_1^{(3)} = -7.77$
				$\hat{\beta}_1^D = 0.0149$	$\hat{\beta}_2^D = 0.985119$	$\hat{\beta}_3^D = 9 \times 10^{-6}$
		(0,26)	(0.0003)	(0.69)	(0.34)	(0.17)

SDMCD	R^2	Moran's I	ρ	Sector 1	Sector 2	Sector 3	Sector 1	Sector 2	Sector 3
	0.74	0.07	0.58	$\gamma_1^{(1)} = -2.00$	$\gamma_1^{(2)} = 11.8$	$\gamma_1^{(3)} = -9.8$	$\theta_1^D = -10.11$	$\theta_2^D = 13.07$	$\theta_3^D = -2.96$
		(0.26)	(0.0003)	(0.56)	(0.11)	(0.09)	(0.02)	(0.21)	(0.74)

- SDMCD provides a better interpretation for the first sector compared to the SEMCD.
- the second and the third parts are positively and negatively connected with *PM*2.5 concentration, respectively.
- The other coefficient θ's indicates how a city's air quality is related to near cities' industry structure.

Conclusion and perspectives

- Many more models could be imagined, mixing functional and compositional predictors
- Taking into account spatial autocorrelation avoids biases and provides safer interpretations
- Overparametrization needs regularization (PCR or PLS)
- Prediction raises specific problems: deleting observations (cross-validation) destroy connectivity [1]
- Variable importance measures could of be interest for interpreting compositional regression analysis [2, 3]

1 Goulard, M., Laurent, T., & Thomas-Agnan, C. (2017). About predictions in spatial autoregressive models: Optimal and almost optimal strategies. *Spatial Economic Analysis*, *12*(2–3), 304–325.

2 Grömping, U. (2015). Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, *7*(2), 137-152.

3 Wallard, H. (2015). Using explained variance allocation to analyse importance of predictors. In *16th ASMDA conference proceedings*, 1043-1054.