Detection and localization of changes in panels of random densities

Piotr Kokoszka

Colorado State university

Work with

Tim Kutta, Colorado State University

Agnieszka Jach, Hanken School of Economics (Helsinki)

Michel Ferreira Cardia Haddad, Queen Mary University of London

Haonan Wang, Colorado State University

Motivation: Brazilian Covid data

Data:

- **What?** Infection data Ct-values
- **Where?** Brazil resolution down to cities
- **When?** 05/15/2020 – 05/15/2022 daily, $\sim$ 105 weeks

Aim:

- Identify changes in space and time
- Changes in viral loads

Data (i)

Typical studies: Counts of cases or deaths

Data (II)

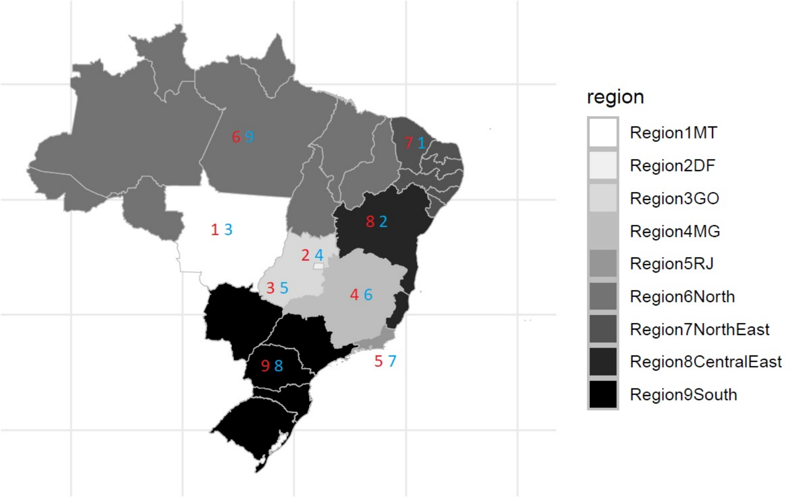
Ct-values:

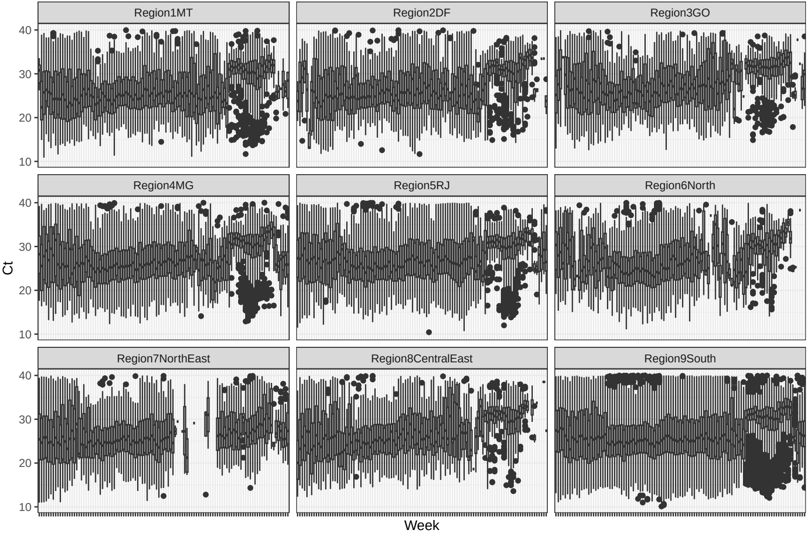
- Measured by PCR test
- Time it takes to replicate measurable amount of virus
- Measure of viral load (between 10 and 40)
- lower number = more virus

Data Aggregation:

- $\approx 300,000$ observations (1.5 million with covariates)
- No or few tests on many days in many regions
- Time aggregation: weeks
- Space Aggregation: 9 regions

Brazil


6 9
7 1
1 3
8 2
2 4
3 5
4 6
9 8
5 7
region
Region1MT
Region2DF
Region3GO
Region4MG
Region5RJ
Region6North
Region7NorthEast
Region8CentralEast
Region9South



Region1MT
Region2DF
Region3GO
Region4MG
Region5RJ
Region6North
Region7NorthEast
Region8CentralEast
Region9South
Ct
Week
10
20
30
40

STATISTICAL MODEL

Indexes:

- $s = 1, ..., S$ for **s**pace
- $t = 1, ..., T$ for **t**ime
- $N_{s,t}$ **n**umber of measurements at (s, t)

Data model:

For all (s, t) $f_{s,t}$ is a random density.

$$X(s, t, i) \sim f_{s,t}, \qquad i = 1, ..., N_{s,t}.$$

Objective:

- Structural breaks in the $f_{s,t}$ (in time and space)

Building blocks

- **The centered log transformation**

$$\Psi : f \mapsto \log[f] - \frac{1}{|I|}\int_I \log[f(x)]dx$$

- Ψ is bijective and takes (positive) densities on I to $L_0^2(I)$
- This defines the Hilbert structure of the **Bayes space**

$$f \oplus g := \Psi^{-1}(\Psi[f] + \Psi[g]), \quad \|f\| := \left\{ \int_I \Psi[f](x)^2 dx \right\}^{1/2}.$$

Change point model:

- $$\Psi[f_{s,t}] = \varepsilon_{s,t} + \begin{cases} \mu_s^{(1)}, & \text{for} \quad t \leq \lfloor T\theta_s \rfloor \\ \mu_s^{(2)}, & \text{for} \quad t > \lfloor T\theta_s \rfloor. \end{cases}$$
- θ_s describes the time of the change
- A change occurs at s when $\mu_s^{(1)} \neq \mu_s^{(2)}$

Aim: Infer set of spatial locations

$$\mathcal{A}^{(changes)} := \{s : \ \mu_s^{(1)} \neq \mu_s^{(2)}\}$$

and times of changes.

Aim of inference

Recall

$$\mathcal{A}^{(changes)} := \{s : \ \mu_s^{(1)} \neq \mu_s^{(2)}\}.$$

Aim:

- Create an estimator $\widehat{\mathcal{A}}$ for $\mathcal{A}^{(changes)}$
- Consistency:

$$\mathbb{P}\big(\mathcal{A}^{(changes)} \subset \widehat{\mathcal{A}}\big) \to 1$$

- Level-α:

$$\mathbb{P}\big(\mathcal{A}^{(changes)} \neq \widehat{\mathcal{A}}\big) \to \alpha.$$

Our tools:

- An asymptotic test $\varphi[\mathcal{A}]$ of "no change in $\mathcal{A}$"

$$H_0 : \mathcal{A} \cap \mathcal{A}^{(changes)} = \emptyset$$

Change point model:

- Recall the data

$$X(s,t,i) \overset{i.i.d.}{\sim} f_{s,t}, \qquad i = 1, ..., N_{s,t}$$

- We consider kernel density estimators KDE

$$\hat{f}_{s,t}(x) := \frac{1}{N_{s,t} h_{s,t}} \sum_{i=1}^{N_{s,t}} K\Big(\frac{x - X(s,t,i)}{h_{s,t}}\Big)$$

- Consider $\Psi[\hat{f}_{s,t}]$ as proxies for $\Psi[f_{s,t}]$.

Sparsity (I)

Sparsity problem:

- $\hat{f}_{s,t} - f_{s,t}$ is not asymptotically negligible.
- $\hat{f}_{s,t}$ can be 0 and $\Psi(\hat{f}_{s,t})$ is not defined
- $\hat{f}_{s,t}$ may not even have the Fréchet mean

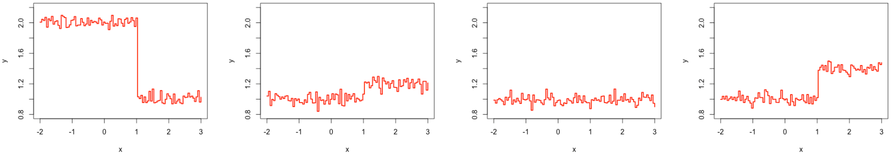
Remedies:

- Restriction to subinterval J, where $f_{s,t} \geq c > 0$.
- "Well-behaved version" $\breve{f}_{s,t}$ of $\hat{f}_{s,t}$ (equal with high probability)

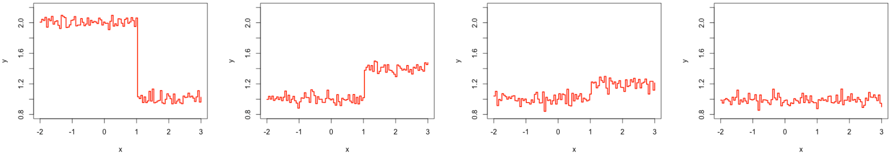
Mathematical details will follow.

SEQUENTIAL TESTING

Unordered locations:



Ordered locations s_1, s_2, s_3, s_4:



Sequential testing: Elimination

Statistically order locations $\hat{s}_1, \hat{s}_2, \hat{s}_3, \hat{s}_4$ from the smallest to largest P-value of a change point test φ applied to each location.

This ordering is equal to the correct ordering s_1, s_2, s_3, s_4 with probability approaching 1.

Sequential testing: Set $\widehat{\mathcal{A}} = \emptyset$

1) $\varphi[s_1, ...s_4]$ (Is there a change in $\{s_1, ..., s_4\}$)?
   - Yes: Update $\widehat{\mathcal{A}} = \{s_1\}$ and move on
   - No: stop
2) $\varphi[s_2, ...s_4]$ (Is there still a change in $\{s_2, ..., s_4\}$)?
   - Yes: Update $\widehat{\mathcal{A}} = \{s_1, s_2\}$
   - No: stop
3) ...

Construction of φ:

- CUSUM statistic

$$\begin{aligned}\widehat{\Delta}[\mathcal{A}] &= \frac{1}{|\mathcal{A}|}\sum_{s\in\mathcal{A}}\frac{1}{T^2}\sum_{t=1}^{T}\Big\|\sum_{r=1}^{t}\Psi[\hat{f}_{s,r}] - \frac{t}{T}\sum_{r=1}^{T}\Psi[\hat{f}_{s,r}]\Big\|^2\\ &\approx \frac{1}{|\mathcal{A}|}\sum_{s\in\mathcal{A}}\frac{1}{T}\sum_{t=1}^{T}\Big\|\frac{1}{\sqrt{T}}\sum_{r=1}^{t}\Psi[\check{f}_{s,r}] - \frac{1}{\sqrt{T}}\frac{t}{T}\sum_{r=1}^{T}\Psi[\check{f}_{s,r}]\Big\|^2\\ &\xrightarrow{d} \frac{1}{|\mathcal{A}|}\sum_{s\in\mathcal{A}}\int_0^1 \|W_s(x) - xW_s(1)\|^2 dx.\end{aligned}$$

- $\{W_s\}_{s=1,\dots,S}$ is a Brownian motion in $\{L_0^2(J)\}^S$.
- $\{W_s\}_{s=1,\dots,S}$ has the same covariance as $\{\Psi(f_{s,1})\}_{s=1,\dots,S}$.

  Spatial dependence estimated, MC to get the distribution of the limit.

Theory (i)

Test decision:

$$\varphi_T[\mathcal{A}] := \begin{cases} 0, & \text{if} \quad \widehat{\Delta}[\mathcal{A}] \leq q_{1-\alpha}[\mathcal{A}], \\ 1, & \text{otherwise.} \end{cases}$$

On a set of probability approaching 1, $\check{\Delta}[\mathcal{A}] = \widehat{\Delta}[\mathcal{A}]$. These two random variables have the same asymptotic distribution.

On a set of probability approaching 1, for each subset of regions, we can detect change with probability approaching 1, if it occurs (consistency of $\varphi_T[\mathcal{A}]$).

We have a fixed (finite) number of regions.

Illustration of the main idea

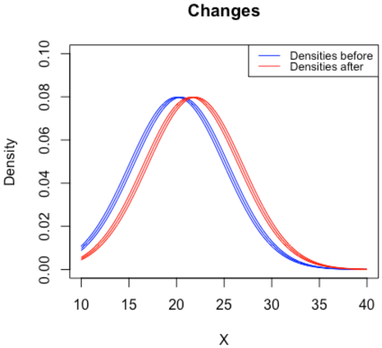
Replace asymptotic probability 1, by probability 1. Suppose $S = 4$, $\mathcal{A}^* = \{1, 2\}$ (change in 1 and 2, but not in 3 or 4).

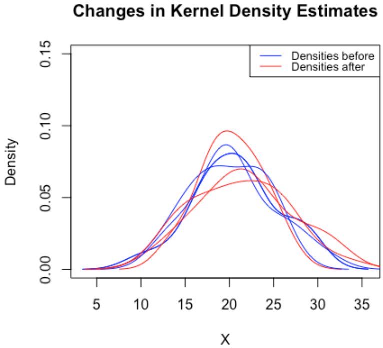
H_0^0: no change in $\{1, 2, 3, 4\}$. Change point detected with probability 1,
$\rightarrow$ test
H_0^1: no change in $\{2, 3, 4\}$. Change point detected with probability 1,
$\rightarrow$ test
H_0^2: no change in $\{3, 4\}$. Error with probability α (change in 3 or 4),
$\rightarrow$ test
H_0^3: no change in $\{4\}$. Error with probability α (change in 4),

$\widehat{\mathcal{A}} = \{1, 2, 3\} \iff$ reject H_0^2 $(\alpha) \rightarrow$ accept H_0^3 $(1 - \alpha)$
$\widehat{\mathcal{A}} = \{1, 2, 3, 4\} \iff$ reject H_0^2 $(\alpha) \rightarrow$ reject H_0^3 (α)

Probability of misidentifying $\mathcal{A}^* = \{1, 2\}$:
$\alpha(1 - \alpha) + \alpha\alpha = \alpha$.

Some simulations





Parameters: $T = 100, N = 50, \alpha = 0.05, S = 5$

	Small change	Large change
1 region	0.95 (0.82)	0.96 (0.82)
2 regions	0.90 (0.85)	0.95 (0.86)
3 regions	0.90 (0.89)	0.96 (0.91)

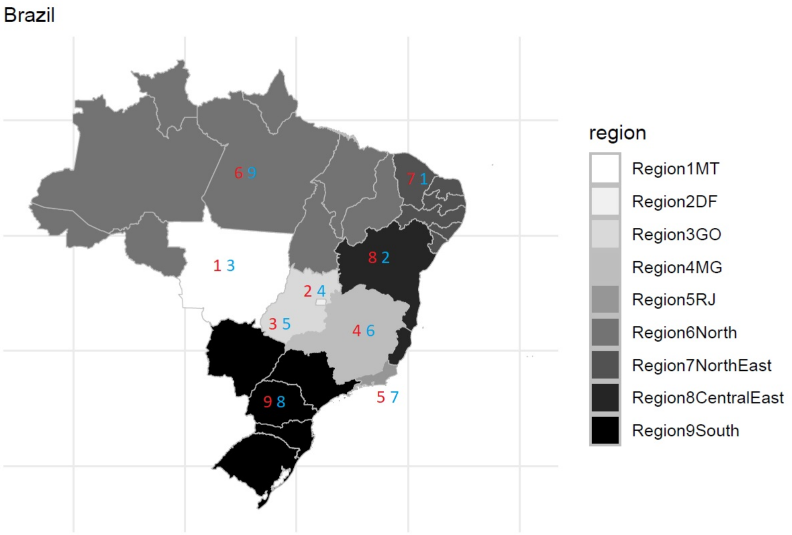
RESULTS (I)

Date	5-11 July of 2021	18-24 October of 2021	1-7 November of 2021
$t*$	62	77	79
year/week	2021/27	2021/42	2021/44
region	Region7NorthEast	Region8CentralEast	Region1MT
Date	1-7 November of 2021	1-7 November of 2021	15-21 November of 2021
$t*$	79	79	81
year/week	2021/44	2021/44	2021/46
region	Region2DF	Region3GO	Region4MG
Date	15-21 of November 2021	15-21 of November 2021	22-28 of November 2021
$t*$	81	81	82
year/week	2021/46	2021/46	2021/47
region	Region5RJ	Region9South	Region6North

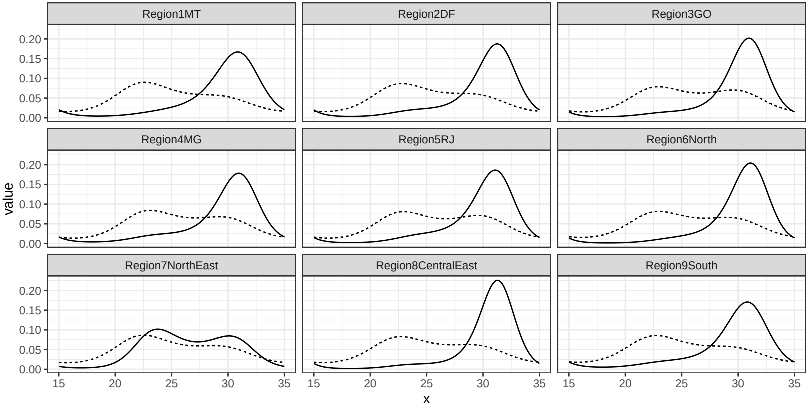
Most changes in November 2021

Delta replaced by omicron in January 2022, with initial reports in November.

RESULTS


Brazil
6 9
7 1
1 3
8 2
2 4
3 5
4 6
9 8
5 7
region
Region1MT
Region2DF
Region3GO
Region4MG
Region5RJ
Region6North
Region7NorthEast
Region8CentralEast
Region9South

RESULTS (II)



Fréchet mean densities before (dashed) and after (continuous) the change points.

Omicron more infectious but less severe.

Theory (II)

Properties of $\breve{f}_{s,t}$: There exists an approximate version $((\breve{f}_{s,t})_s)_{t=1,\dots,T}$ of $((\hat{f}_{s,t})_s)_{t=1,\dots,T}$ such that

$$\mathbb{P}\big(\breve{f}_{s,t}(x) = \hat{f}_{s,t}(x),\ \forall x \in J,\ \forall s,t\big) \to 1, \quad as \quad T \to \infty, \qquad (1)$$

with $((\breve{f}_{s,t})_s)_{t=1,\dots,T}$ independent across t. The version satisfies for some fixed $c > 0$ the moment condition

$$\sup_{s,t} \mathbb{E}\|\Psi[\breve{f}_{s,t}]\|^4 \leq c \qquad (2)$$

and the mean approximation property

$$\|\mathbb{E}\Psi[\breve{f}_{s,t}] - \mathbb{E}\Psi[f_{s,t}]\| = o(T^{-1/2}). \qquad (3)$$

Theory (III)

$$\breve{G}_T(x) := \Big(\frac{1}{\sqrt{T}} \sum_{r=1}^{\lfloor xT \rfloor} \breve{Y}_{s,r} - \mathbb{E}[\breve{Y}_{s,r}]\Big)_s, \qquad x \in [0,1],$$

The $\breve{Y}_{s,r} = \Psi(\breve{f}_{s,t})$ are not identically distributed.

Proposition: Under our assumptions,

$$\{\breve{G}_T(x)\}_{x\in[0,1]} \xrightarrow{d} \{W(x)\}_{x\in[0,1]}$$

in the $J1$ topology of the Hilbert space $(L_0^2(J))^S$.