BANQUE DE FRANCE
EUROSYSTÈME

CAN WE PREDICT INDUSTRIAL PRODUCTION USING SATELLITE DATA?

NON-TRADITIONAL DATA AND STATISTICAL LEARNING WITH APPLICATIONS TO MACROECONOMICS

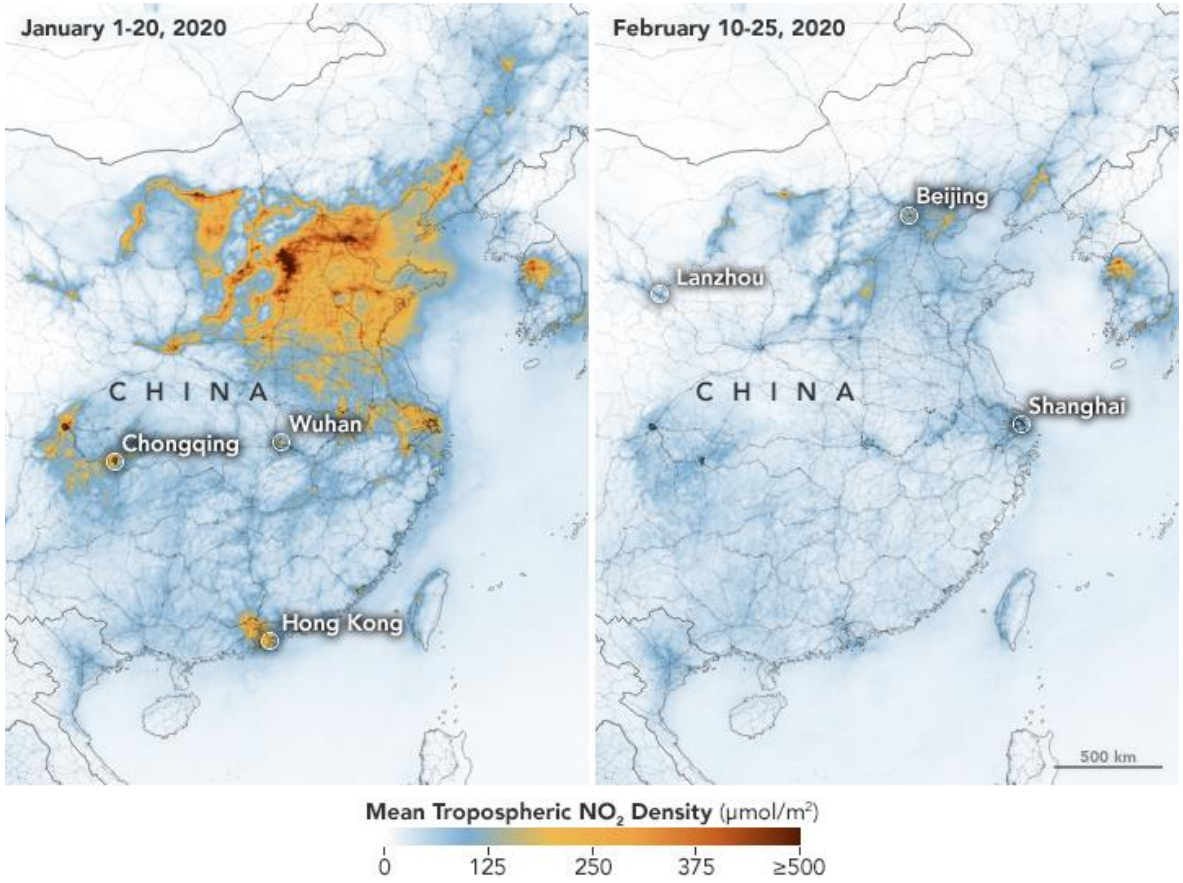
NOVEMBER 12TH

J-C. BRICONGNE[1], B. MEUNIER[1], T. PICAL[2]

(1) BANQUE DE FRANCE ; (2) AIX-MARSEILLE SCHOOL OF ECONOMICS

FULL SCRIPT AND DATASET AVAILABLE ON GITHUB.COM/THOMASPICAL/SENTINEL5_NO2

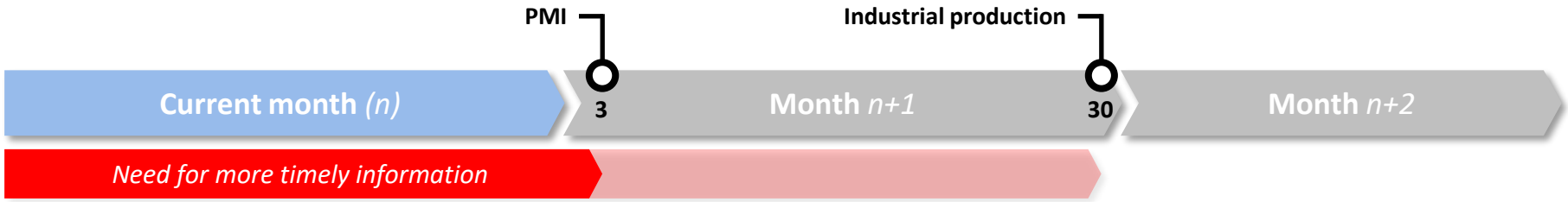
MOTIVATION



*Can NO_2 pollution data help **predict industrial production** ?*

WHY TURNING TOWARDS SATELLITE DATA ?

1. **Timeliness**: daily data available the following day



Timeline example for Japan – March 2020

2. Extensive **geographic coverage** – including over developing countries in which industrial production statistics and/or alternative pollution data are scarce

3. **Uniqueness** of the sensor – limiting the risk of idiosyncratic particularities/errors when pooling together multiple sensors

4. **Uniform coverage** – no composition effect due to the arbitrary location of ground sensors across a region

DATA SOURCE

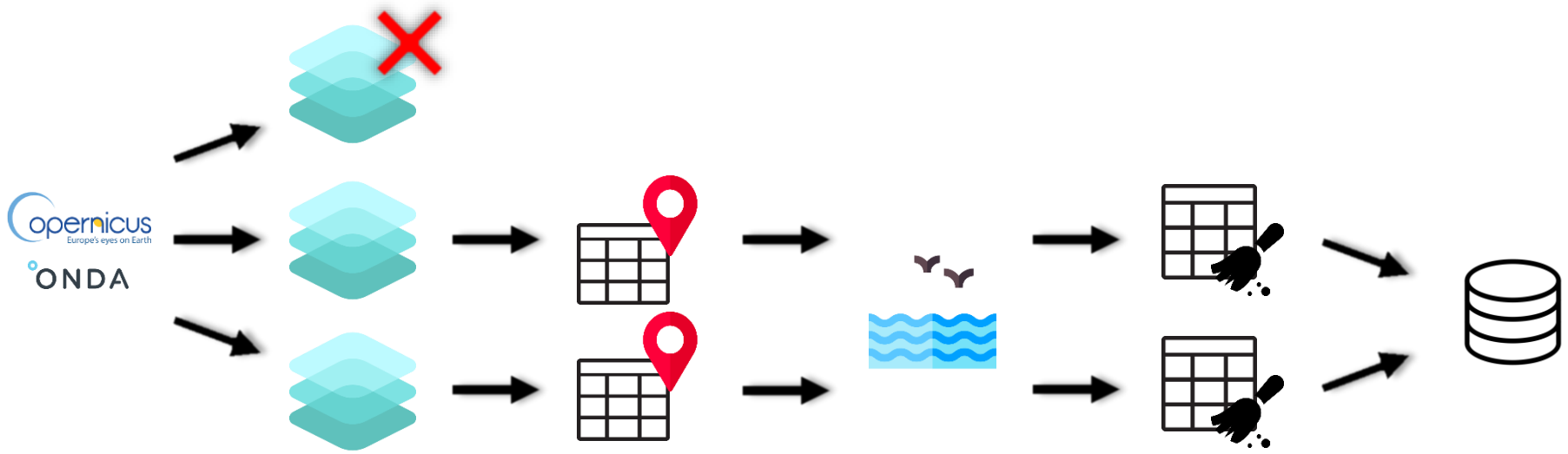
- **Satellite data** from the European Spatial Agency's Sentinel 5P: one observation per day for each 7x3.5 km point of Earth

- **What we got is a concentration of NO_2** (nitrogen dioxide) in the troposphere:
  - Gas emitted by the **combustion of fossil fuels** in industrial production, transportation and energy production
  - **Early pollutant** since this molecule is a precursor of other pollutants

- On top of common challenges for high-frequency data, data quality is affected by **clouds and snow**

- **Literature** linking **economic growth** and **pollution**, use of a pollution as a high frequency tracker of economic activity, but no literature using NO2 pollution to forecast economic variables. (*detailed literature in appendix*)

RETRIEVING AND TREATING DATA



Missing NO_2 data

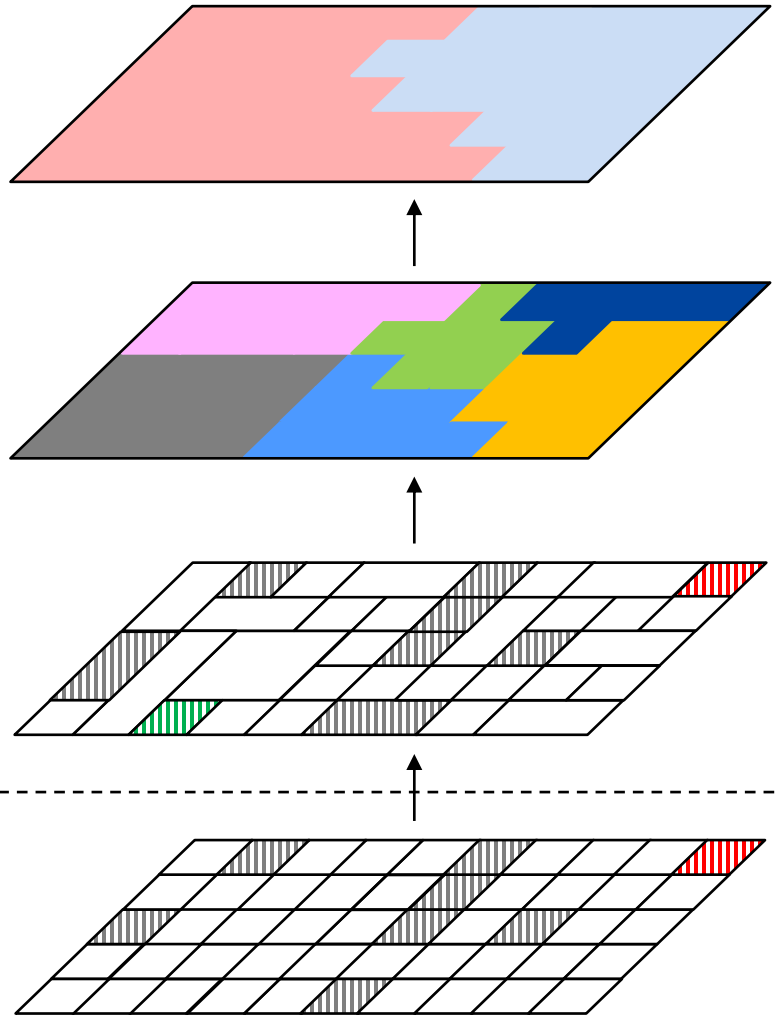
Quality > 0.75

Files download and check | **Flattening data** | **Ocean detection** | **Cleaning data (quality index)** | **Aggregating data**

Data size — *3-4 Go per day* → *10 Mo per day*

BANQUE DE FRANCE
EUROSYSTÈME

AGGREGATING AND CORRECTING DATA



8 *Overweight urban / industrial areas*

7 *Correct for meteorological factors (He et al., 2017)*

6 *Correct for seasonality (Ollech, 2018)*

5 *Select optimal framing by cross validation*

4 *Interpolate missing data*

3 *Eliminate outliers*

2 *Drop ZIP code if too many missing constituents*

Aggregation and quality check already performed.

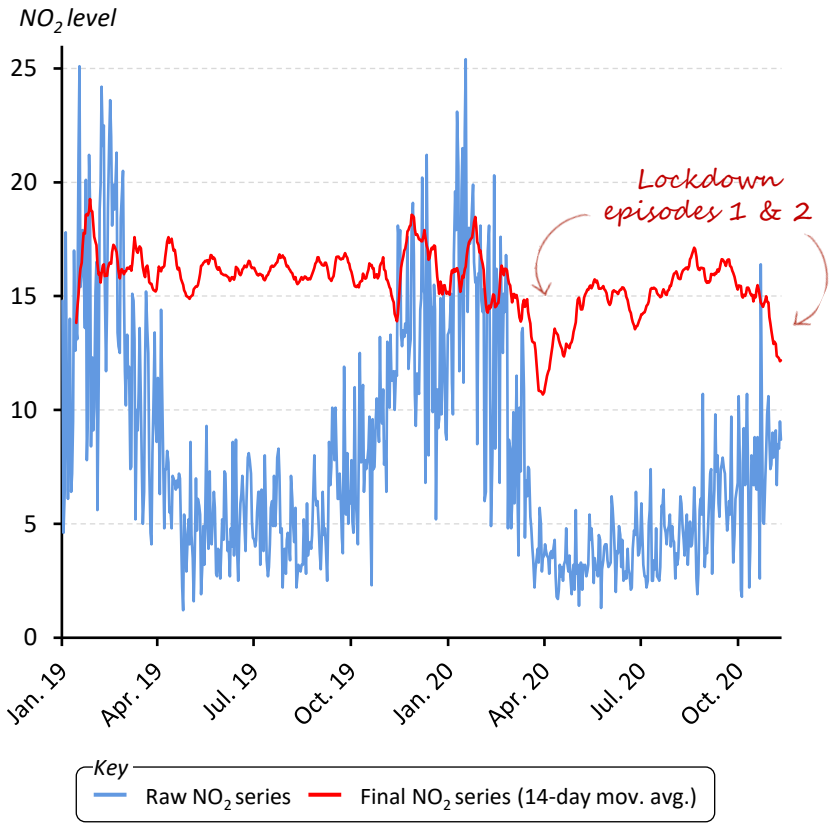
1 *Eliminate data if quality < 0.75*

Types of missing points
Missing NO_2 data | Quality < 0.75 | Outlier

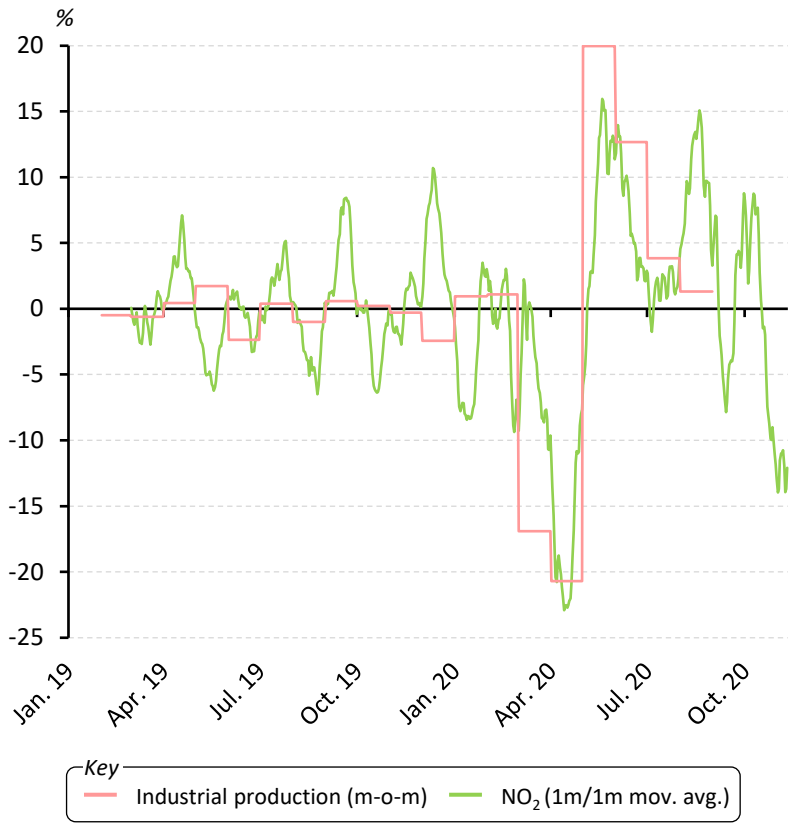
BANQUE DE FRANCE
EUROSYSTÈME

RESULTING SERIES

Area-level
(e.g. Grenoble, France)



Country-level
(e.g. France)



BANQUE DE FRANCE
EUROSYSTÈME

ECONOMETRIC MODELLING

- Stationarity of high-frequency data: year-on-year difference might work (as in Ferrara and Simoni, 2019 or Lewis *et al.*, 2020) but introduce a **base shift** and potential spurious cycle (Ladiray *et al.*, 2018)

  → Instead rely on a **moving average over one month over the previous month**

- Accommodate for mixed-frequencies data: monthly industrial production and daily pollution

  → Build on **MIDAS model class** (Ghysels *et al.*, 2004) which allows to
  - Put different weight on different lags – first weeks might be more important for month-on-month growth, or last data points might be more crucial
  - Bring lots of lags while preserving parsimony – in cases where weights follow a distribution function independent of the number of lags (e.g. Almon polynomial, exponential)

- Limited time sample for a single country since Sentinel-5P was launched only on end-2018

  → Rely on **panel estimates** since data is available for all countries – ultimately would combine panel and MIDAS using Khalaf *et al.* (*Journal of Econometrics*, in press)'s panel-MIDAS framework

RESULTS

Procedure

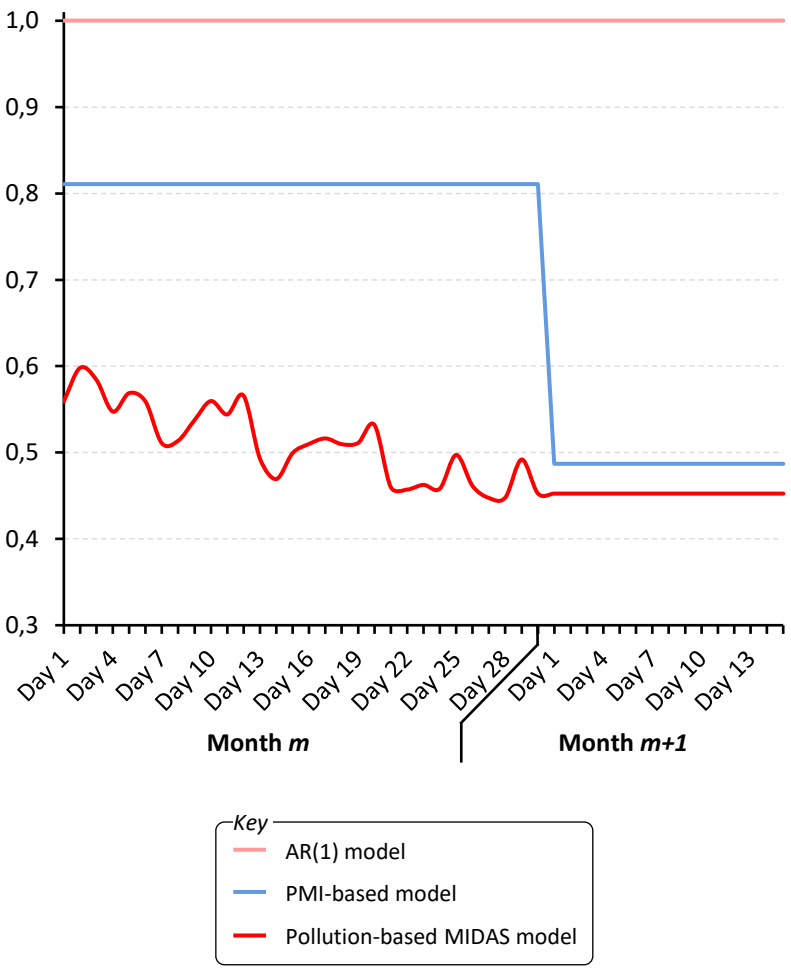
- Compare across **three different models**: AR(1), PMI-based, and MIDAS with daily satellite data
- **One-period ahead out-of-sample forecast** ; first model estimated up to March 2020 then expanding window by one month each step
- **Real-time** exercise with data available at the end of each day

Results

- **Lower out-of-sample RMSE** for MIDAS model until the start of the following month *m+1* when PMIs become available ; still outperform but difference becomes non-significant
- RMSE on a **decreasing trend** during the month as more data for the month *m* becomes available

Out-of-sample RMSE relative to the AR(1) model

1=AR(1) model performance



BANQUE DE FRANCE
EUROSYSTÈME

CONCLUSIONS AND NEXT STEPS

1. **Work still in progress**. Satellite are promising but need to be corrected for **several specific issues** (data quality, aggregation level, meteorological factors) **on top on of common statistical issues pertaining to high-frequency data** (multiple levels of seasonality, noise).

2. The **signalling power** of satellite data swiftly detects turning points in industrial activity during crisis – however too early to conclude on nowcasting performances during "normal" times (high-frequency data might be of second-order when economic conditions are stable)

3. Defining an appropriately large zoning is key to track pollution adequately – therefore this indicator might be limited for **small countries** as pollution spill over borders

4. Satellite data might provide valuable information for **developing countries** where official statistics for industrial activity are limited – might be used to develop a PMI-like indicator of manufacturing activity based on NO_2 pollution

5. Other types of satellite data exist and might be exploited – e.g. **infrared data** detecting heat produced by factories

BANQUE DE FRANCE
EUROSYSTÈME

APPENDIX

BANQUE DE FRANCE
EUROSYSTÈME

FULL SCRIPT AVAILABLE ON GITHUB:

HTTPS://GITHUB.COM/THOMASPICAL/SENTINEL5_NO2

SERVICES AND TOOLS USED

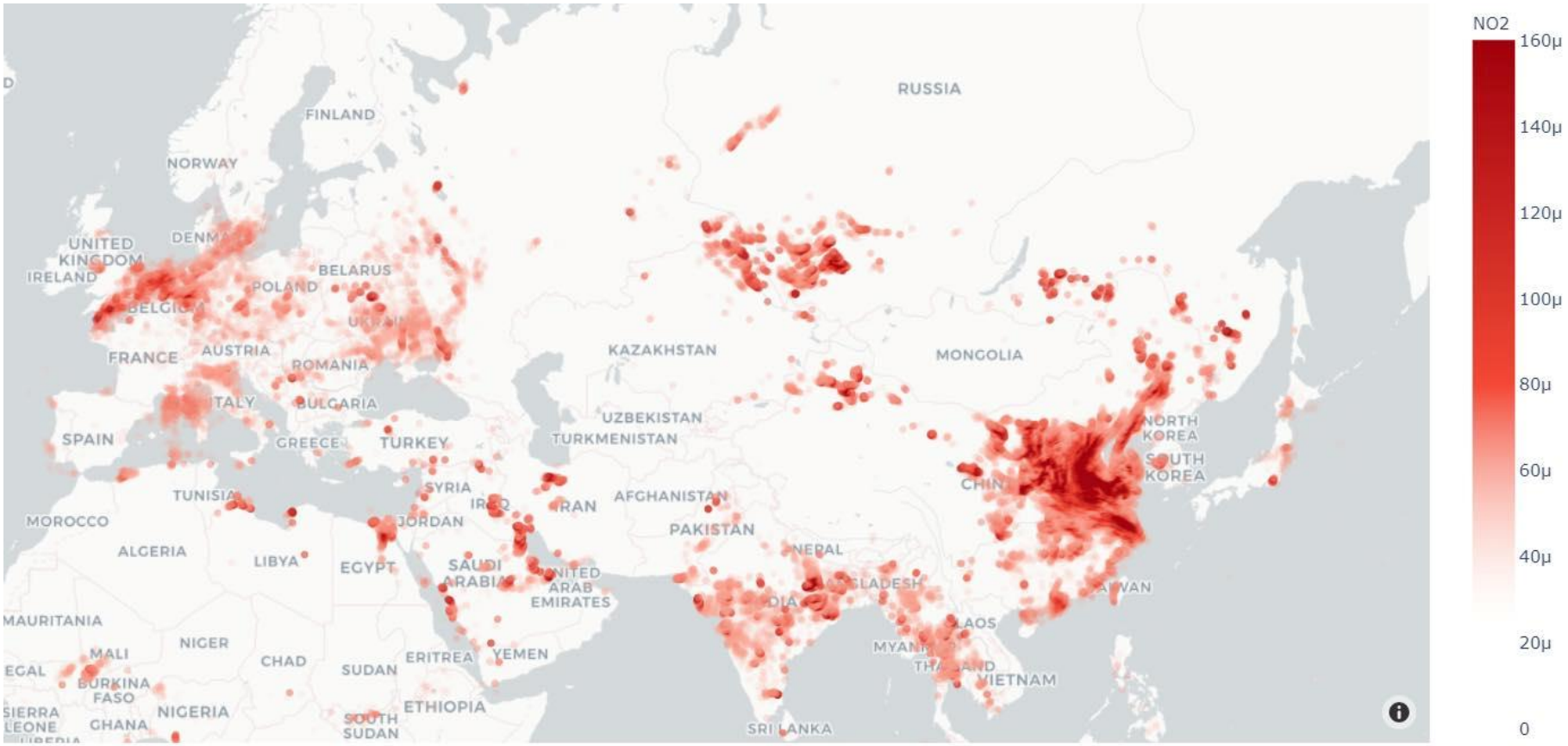
Data provider

Data script

Data visualisation

SENTINEL 5P FUNCTIONING

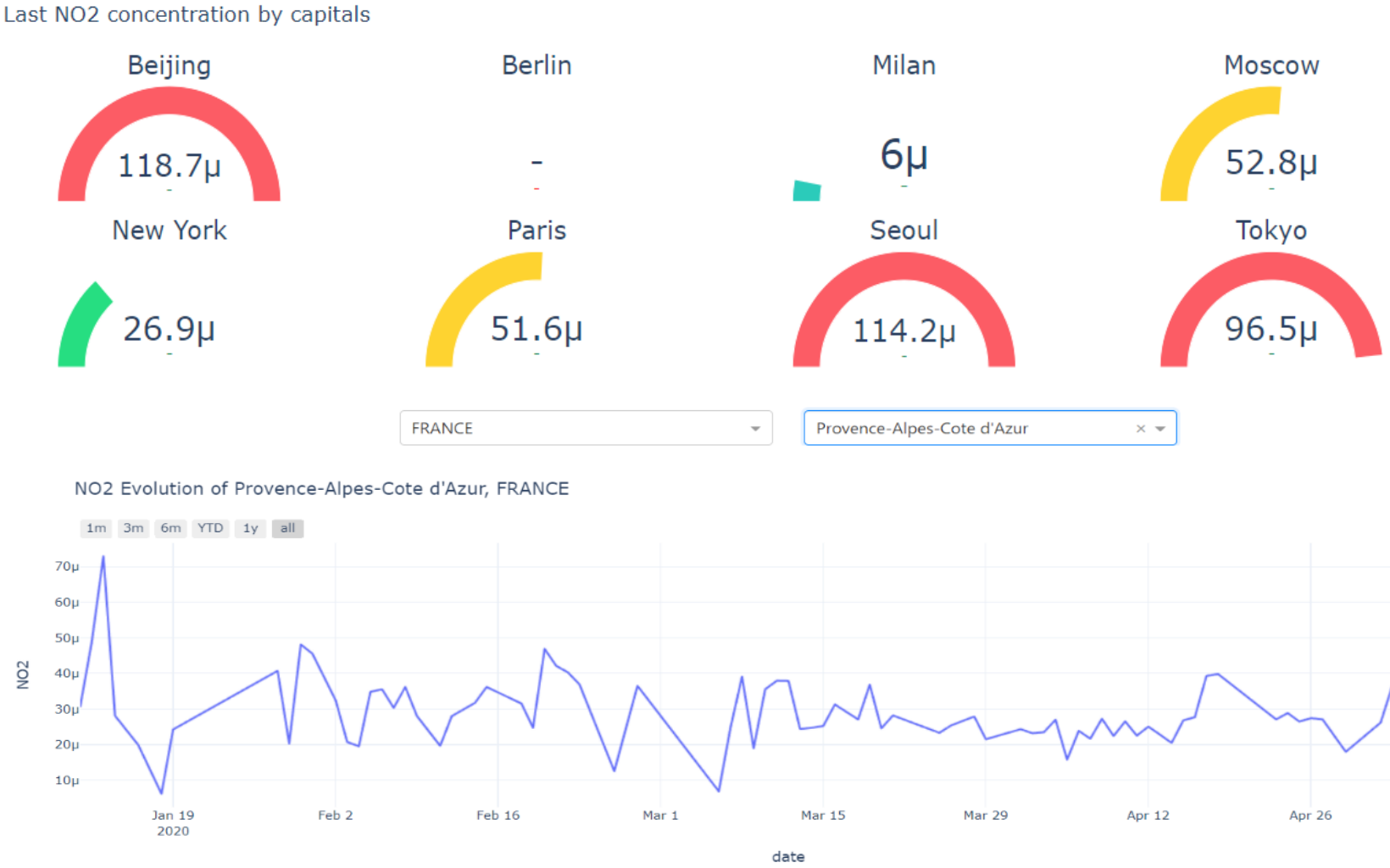
VISUALISATION OF RAW DATA



DATA EXAMPLE

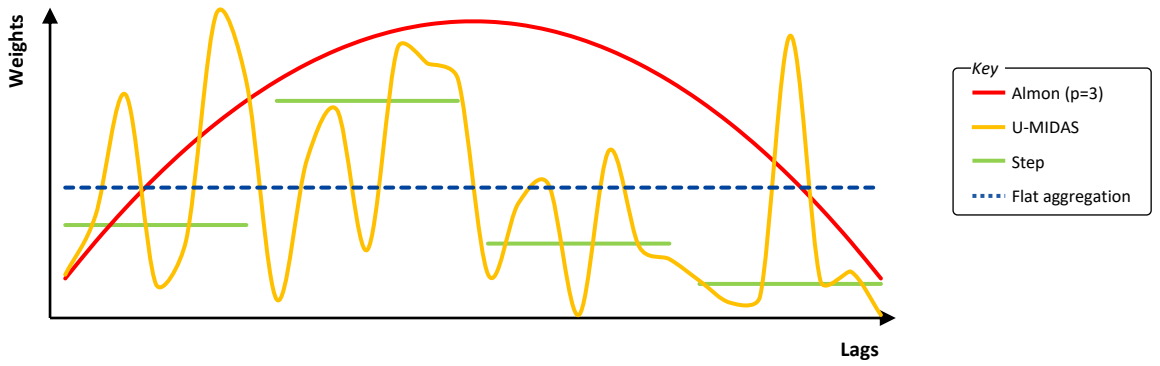
year	month	week	cc_pays	cc_departement	cc_region	cc_ville	longitude	latitude	NO2	quality	hour_mean	hour_std	dayofweek_mean	dayofweek_std	day_mean	day_std	counter
2020	7	28	AD	Undefined	Andorra la Vella	Andorra la Vella	1.52	42.5	1.32e-05	1.0	11.0		1		7		1
2020	7	28	AD	Undefined	Canillo	Canillo	1.58	42.6	1.46e-05	1.0	11.0	0.0	10.0		70.0		2
2020	7	28	AD	Undefined	Canillo	El Tarter	1.67	42.6	1.34e-05	1.0	12.2	1.032	10.0		70.0		10
2020	7	28	AD	Undefined	Encamp	Encamp	1.60	42.5	1.26e-05	1.0	12.0	1.154	10.0		70.0		4
2020	7	28	AD	Undefined	Encamp	Pas de la Casa	1.76	42.6	1.31e-05	1.0	12.2	1.032	10.0		70.0		10
2020	7	28	AD	Undefined	La Massana	Arinsal	1.42	42.7	1.48e-05	1.0	12.15	1.014	10.0		70.0		19
2020	7	28	AD	Undefined	La Massana	la Massana	1.49	42.5	2.09e-05	1.0	12.0	1.414	10.0		70.0		2
2020	7	28	AD	Undefined	Ordino	Ordino	1.55	42.6	1.39e-05	1.0	12.33	1.154	10.0		70.0		3

FORMATTING RESULTS (PLATFORM PROVIDED BY AMSE)



MIDAS WEIGHING FUNCTIONS

	Function	Number of parameters
Polynomial of degree p (*Almon*)	$g(f_t, \theta, K, p) = \sum_{k=0}^{K} c(k, \theta) \cdot f_{t-k} \; ; \; c(k, \theta) = \sum_{j=0}^{p-1} k^j \cdot \theta_j$	• 2p
Unrestricted (U-MIDAS)	$g(f_t, \theta, K, J) = \sum_{k=0}^{K} \theta_k \cdot f_{t-k}$	• K
Step	$g(f_t, \theta, K, J) = \sum_{k=0}^{K} \theta_{floor(k/J)} \cdot f_{t-k}$	• ceil(K/J)



BANQUE DE FRANCE
EUROSYSTÈME

ECONOMETRIC FRAMEWORK

- Modelling step relying on a **panel-MIDAS approach** developed by Khalaf *et al.* (2020, *Journal of Econometrics*)

Rationale for a **panel** approach

- Data across a **lot of countries** (virtually the entire world)
- Limited **number of observations** for a single country (series start end-2018)

Rationale for a **MIDAS** set-up

- Can accommodate for **mixed-frequencies data**: monthly industrial production and daily pollution
- Adjust for **different weights on different lags** – first weeks might more important for m-o-m growth rate
- Can bring **lots of lags while preserving parsimony** if weights follow a particular distribution (polynomial, exponential)

DATA SOURCE AND RELATED LITERATURE

Data sources for NO_2 pollution

- **Satellite data** from the European Spatial Agency's Sentinel 5P: one observation per day for each 7x3.5 km point of Earth
- **Concentration of NO_2** (nitrogen dioxide) in the troposphere:
  - Gas emitted by the **combustion of fossil fuels** in industrial production, transportation and energy production
  - **Early pollutant** since this molecule is a precursor of other pollutants
- On top of common challenges for high-frequency data, data quality is affected by **clouds and snow**

NO_2 pollution in the literature

- Evidence that economic growth increase NO_2 pollution and conversely that **economic crisis do translate into a drop of NO_2 emissions**
  - Great Financial Crisis: Boersma and Castellanos (2012) for EU; Du and Xie (2017) for China
  - Great Lockdown: Filonchyk *et al.* (2020) for China; Diamond and Wood (2020)
- Utilization as **high-frequency indicators** for economic activity in recent papers such as Deb *et al.* (2020) or Bricongne *et al.* (2020)
- But limited use in **forecasting** – in particular for industrial production as PMIs are already an early indicator with high predictive power (INSEE, 2020)