BANQUE DE FRANCE
EUROSYSTÈME

CAN SATELLITE DATA PREDICT INDUSTRIAL PRODUCTION ?

13 MAY 2022

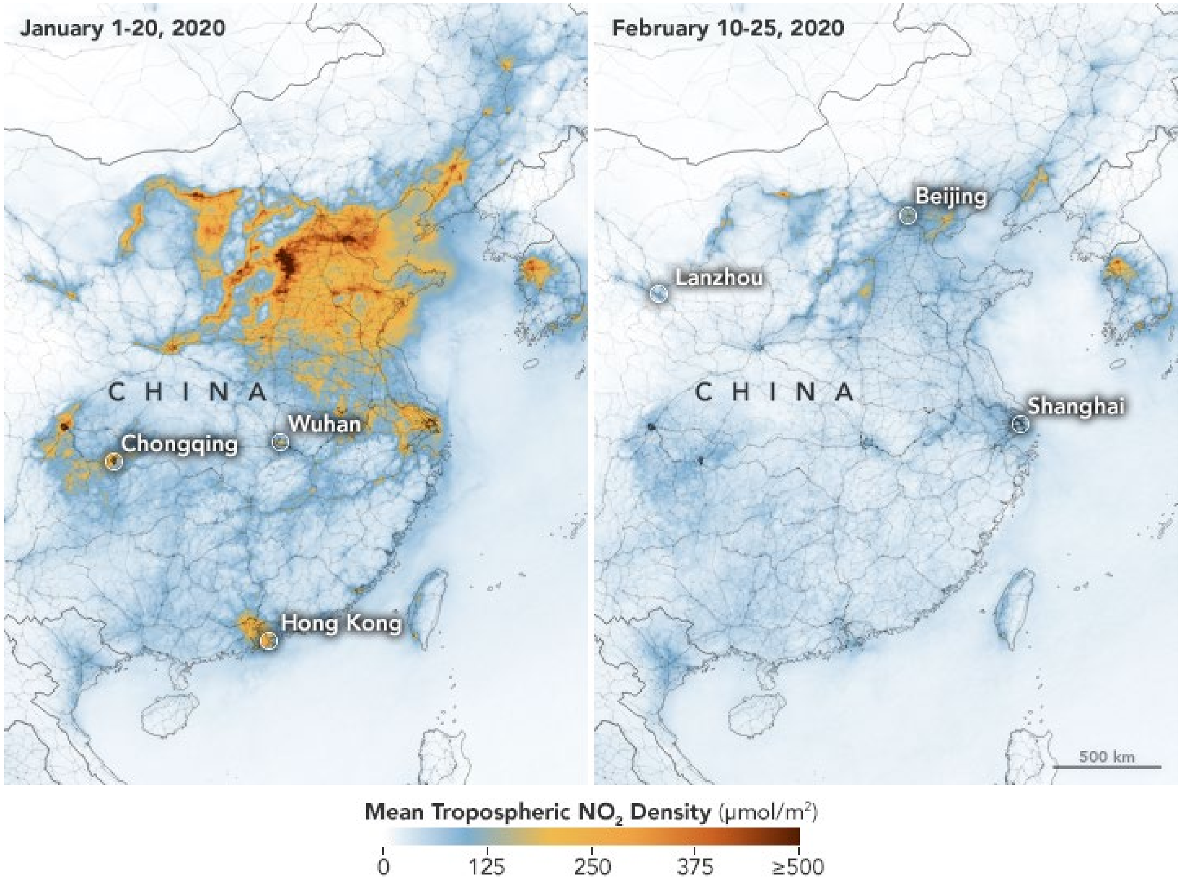
J-C. BRICONGNE[1], B. MEUNIER[1,2], T. PICAL[3]

(1) BANQUE DE FRANCE ; (2) EUROPEAN CENTRAL BANK ; (3) EQUANCY

PAPER AVAILABLE ON [HTTPS://PAPERS.SSRN.COM/ABSTRACT_ID=3967146](https://papers.ssrn.com/abstract_id=3967146)

SCRIPT AVAILABLE ON [HTTPS://GITHUB.COM/THOMASPICAL/SENTINEL5_NO2](https://github.com/thomaspical/sentinel5_no2)

MOTIVATIONS
COVID-19 FROM SPACE



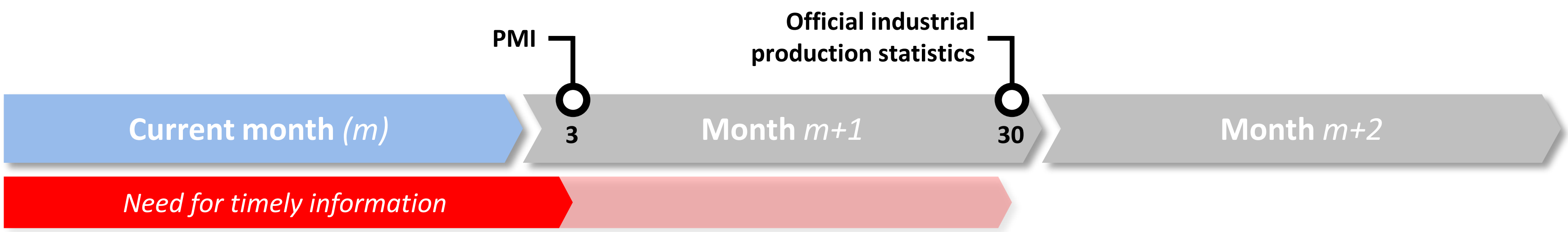
Can NO_2 pollution data help ***predict industrial production*** *?*

MOTIVATIONS
GENERAL APPROACH

Main idea

Forecast in real-time (nowcast) **industrial production** using satellite data on pollution by exploiting:

1. **Timeliness** of satellite data (daily and available on the following day)



Timeline example for Japan – March 2020

2. Features of **nitrogen dioxide** (NO_2):
   - **Earliness:** precursor for other pollutants
   - Emitted by **combustion of fossil fuels** (industrial activity, transportation, coal-fired energy)
   - Low **duration** in the atmosphere

DATA DESCRIPTION

- Data from the **satellite Sentinel 5P**:
  - Launched in 2017 by the European Spatial Agency (ESA) and TROPOMI instrument started in **2018**
  - **Sun-synchronous orbit**: daily passing is at (approximately) the same time every day and at every point (around 13:35 local Sun time)

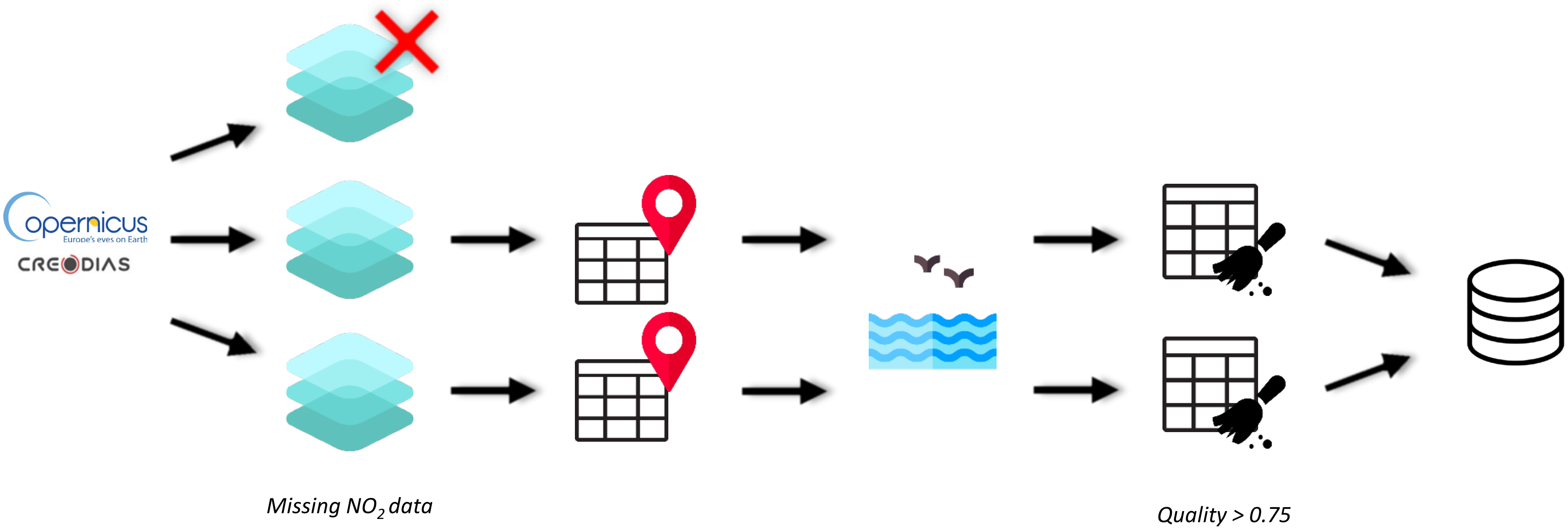
- One observation **per day** for each point of Earth (7 × 3.5 km^2)
- **Concentration of NO_2** in troposphere (lower layer of the atmosphere, 0-15 km)
- Data quality affected by **clouds and snow**; index of data quality between 0 and 1 for each observation

DATA
WHY TURNING TOWARDS SATELLITE DATA FROM THE ESA ?

1. *Vs.* other high-frequency data:
   - Large **geographic coverage** (incl. over developing countries) and **homogenous quality** – in contrast with other data, even those with wide-spectrum (e.g. Google mobility)
   - **Uniqueness** of sensor – no risk of idiosyncratic errors (e.g. if pooling together multiple sensors)
   - **Uniform coverage** – no composition effects (e.g. due to arbitrary location of sensors) or selection bias (e.g. if data retrieved from only certain users)

2. *Vs.* other satellite data on NO_2 pollution:
   - **Measurement errors** documented in the literature for NASA's data (Wang *et al.*, 2020)
   - **Unprecedented precision** of ESA's data with 7×3.5 km² points

DATA
RETRIEVING PROCESS

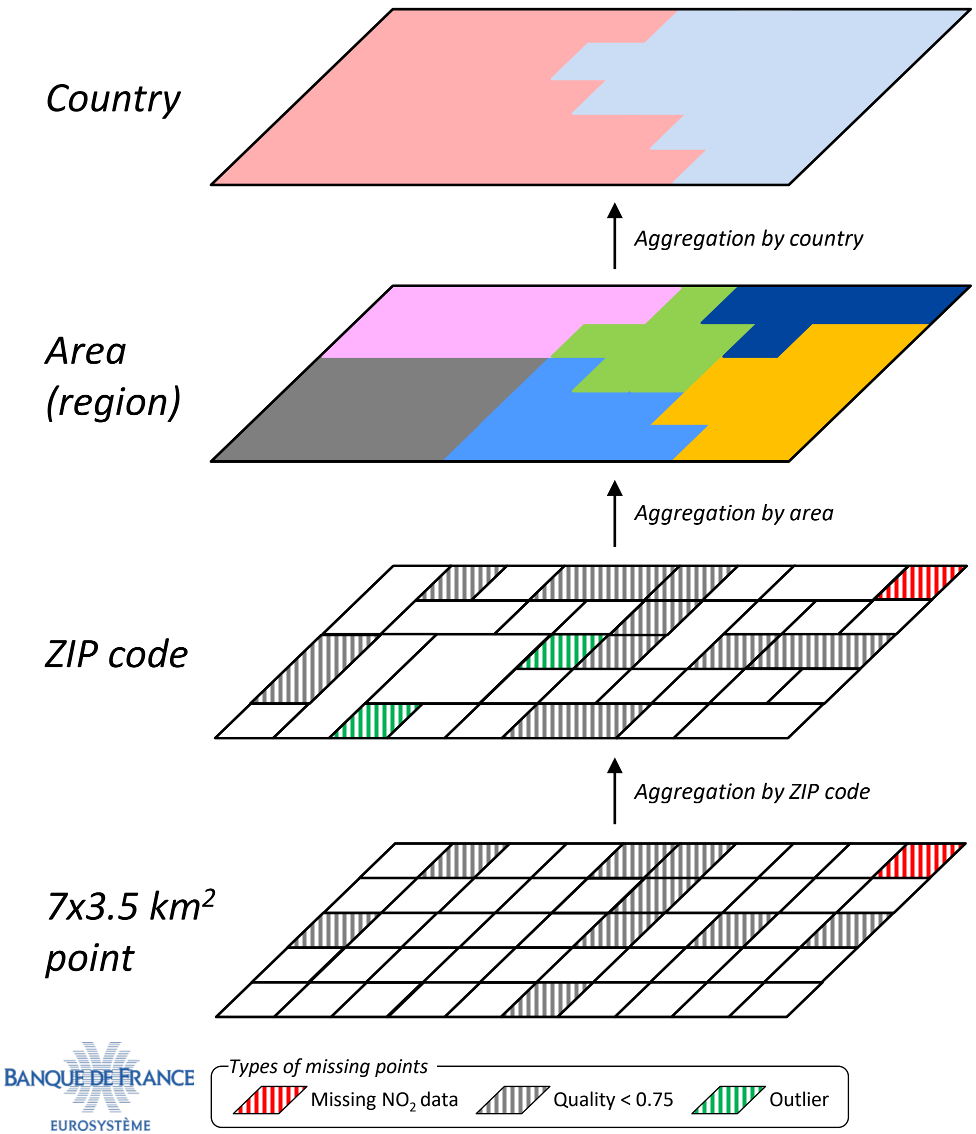


Downloading and checking	Flattening data	Detecting oceans	Cleaning data	Aggregating by ZIP code

Data size	*3-4 Gb per day*	→	*10-20 Mb per day*

BANQUE DE FRANCE
EUROSYSTÈME

DATA
AGGREGATIONS AND ADJUSTMENTS



6. *Correct for weather factors (random forest)*
5. *Interpolate missing data (k-nearest neighbours)*
4. *Eliminate outliers*
3. *Drop ZIP code if too many missing constituents*
2. *Eliminate data if quality < 0.75*
1. *Eliminate oceans (> 30 km from coasts)*

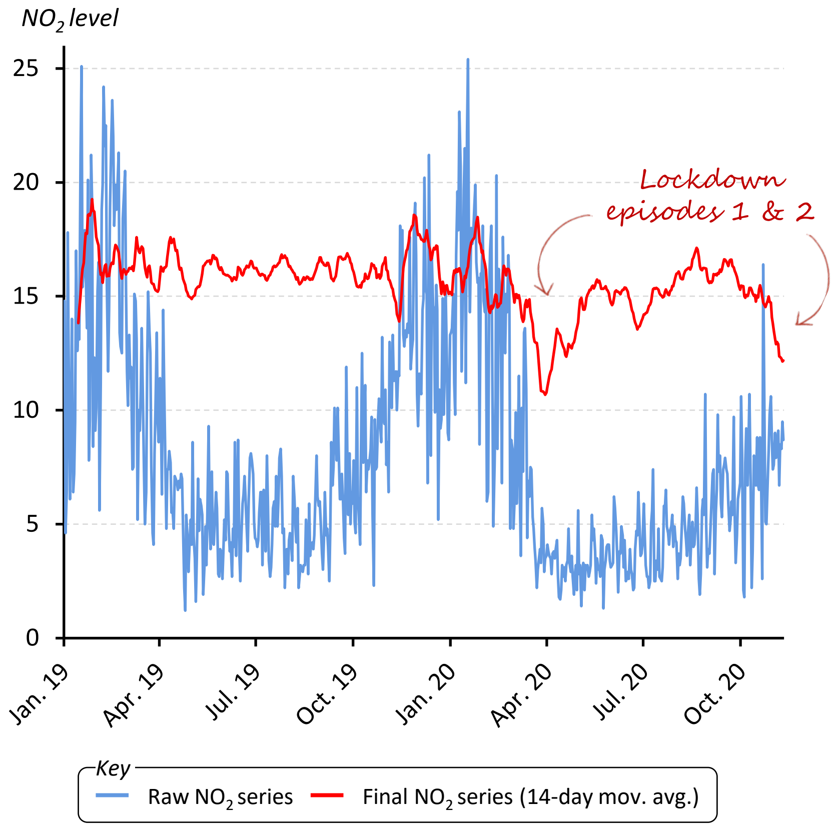
BANQUE DE FRANCE
EUROSYSTÈME

DATA
MACHINE LEARNING TECHNIQUES FOR DATA CORRECTION

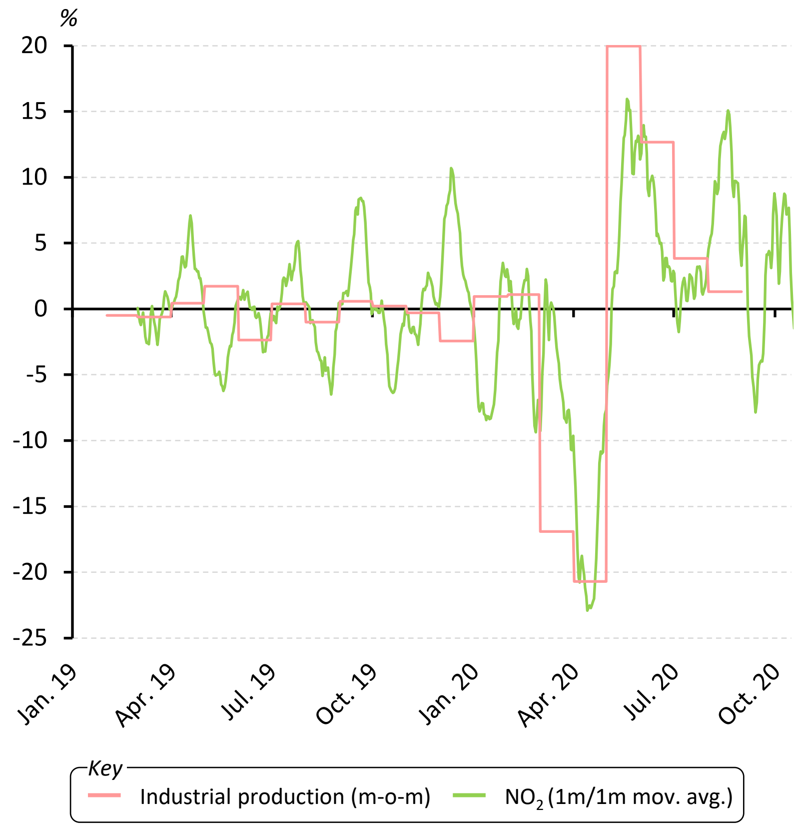
	5 Interpolation of missing points	6 Weather normalization
Initial issue	• **Up to 50% of missing data** not missing at random • Might result in **composition effects**	• Air pollution highly **sensitive to weather** (chemical process, human behaviours) • Effects of weather **greater** than those of policies or economic events (Anh et al., 1997)
In the literature	• Full strand in **geostatistical sciences** with different approaches • Most performant taking **both spatial and temporal** correlations (Yang and Hua, 2018) • But **high computational cost** of most sophisticated algorithms (Kianan *et al.*, 2021) not suited for daily and global data	• Also full strand in **geostatistical sciences** • **Non-linear impact** of weather variables including also interactions between them (Grange et al., 2018) • To be done in **homogenous area** to account for differences in topology and weather (Liu *et al.*, 2020)
In our paper	• **K-Nearest Neighbours algorithm** (KNN) using both time and space (as in Poloczek *et al.*, 2014) to limit computational cost	• Run the weather-normalization **by region** • **Random forest**: non-linearities, low cost, and limited sensitivity to hyper-parameters

DATA
RESULTING SERIES

Area-level
(e.g. Grenoble's region, France)



Country-level
(e.g. France)



NOWCASTING
GENERAL APPROACH

1. **Out-of-sample nowcasting of industrial production growth (month-on-month)**: estimation up to preceding month, then out-of-sample prediction for the current month

2. **Expanding window**: first nowcast for March 2020, then add month by month up to Dec. 2020

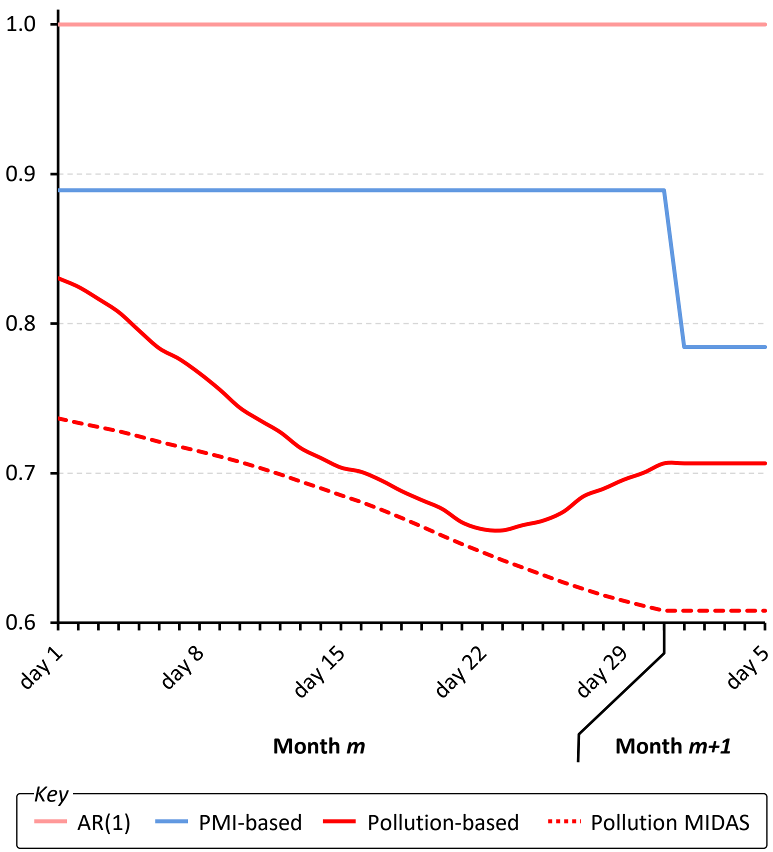
3. Compare performances of pollution-based model with **two benchmarks**
    - **AR(1)** model
    - **PMI-based** model following the literature (Bruno and Lupi, 2003; Tsuchiya, 2014, Akdag *et al.*, 2020)

4. **Daily real-time**: one forecast for each day of the month using the data that would have been available to the forecaster – NO_2 up to preceding day and PMI up to preceding month

5. Rely on **panel estimates** to make up for limited timespan

NOWCASTING
COMPARISON OF PERFORMANCES

Out-of-sample RMSE[1]
Relative to the AR(1) model (=1)



Results

- **Lower out-of-sample RMSE** for pollution-based model
- **Decreasing RMSE** over the month as more information is available
- Performances deteriorate after **day 24:**
  - → Might be due to **first days of month** having more importance for month-on-month growth
  - → Rather rely on the **panel MIDAS** – from Khalaf *et al.* (2021)
  - ⇨ No deterioration at end of the month and **improved performances *vs.* simple averaging**

Note: (1) Panel on 33 countries for which monthly industrial production statistics and PMIs are available (16 advanced and 17 emerging economies)

NOWCASTING
HETEROGENEITIES

	"Crisis" vs. "normal" periods	Share of manufacturing
Initial issue	• High-frequency data might provide useful timely signal during **"crisis"** but might be of second order during **"normal"** periods	• Relevance of NO_2 for nowcasting can depend on importance of **polluting activities** in GVA • Can be linked to **development level** as in Hu and Yao (2019) for "night lights"
Empirical tests	• **Short timespan** does not allow to break by "crisis" and "normal" periods • Instead, break sample depending on the **fall** in industrial production during Covid-19	• Interact the variable of interest with the **share of manufacturing** in value added • Also introduce a **triple interaction** with a dummy for emerging or developing economy
Results	• Greater accuracy gains for the pollution-based model in the countries with **larger drops** in industrial production • Pollution-based model **out-performing benchmarks** (PMI-based) for all sub-samples	• **Share of manufacturing significant**: the higher share of manufacturing the higher the elasticity of pollution to industrial production • Dummy for emerging and developing economies **not significant**

NOWCASTING
REAL-TIME DETECTION OF TURNING POINTS

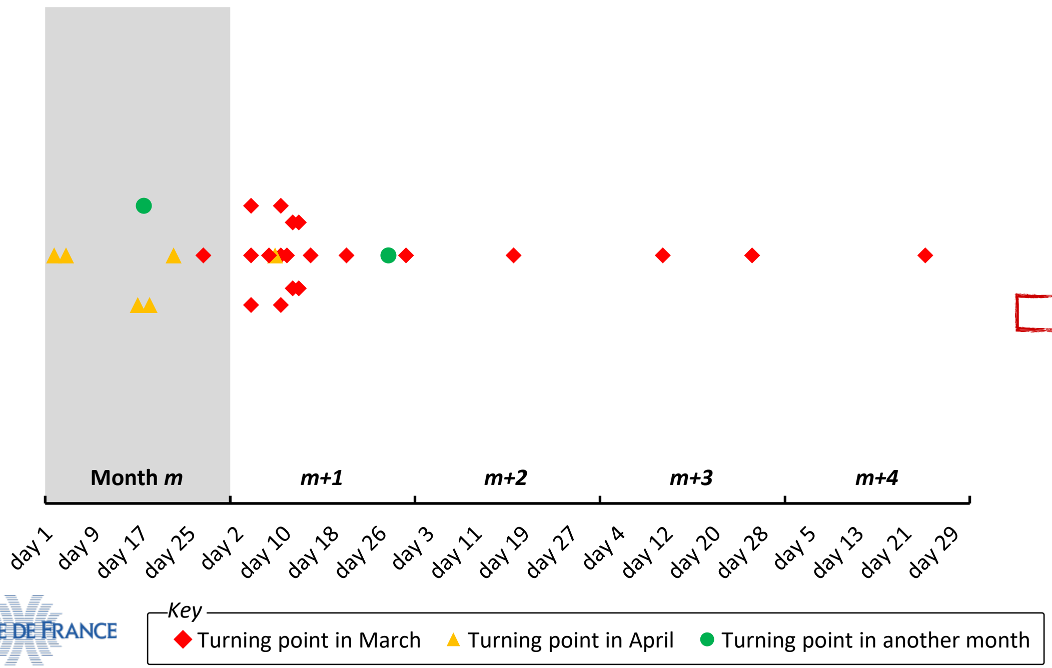
Rationale / procedure

- Daily data might serve for **swifter detection of turning points**

⇨ Following Hamilton (1989), use a **univariate 2-states Markov-switching model** to detect breaks in the time series for a country – a transition to state 2 signals a turning point

⇨ To minimize “false positive”, set **a number of K consecutive periods** in which the MS model has to stay in state 2 before detecting the turning point (empirically, K=21)

Dates for real-time detection of turning points



⇨ Detection generally **1.5 month after Covid-19 outbreak** vs. a 3.5 months delay if using official monthly data

BANQUE DE FRANCE
EUROSYSTÈME

CONCLUSION
MAIN FINDINGS

1. Satellite data to be corrected for specific factors (data quality, missing points, weather) but bring value-added: **timeliness**, **global and uniform coverage**, **granularity**, and **free-to-use**

2. Pollution-based model **strongly out-performs benchmark models** in nowcasting industrial production, with evidence for heterogeneities:
   - **Greater accuracy gains** during larger “crisis” episodes
   - Higher elasticity of NO_2 pollution to industrial production **if higher share of manufacturing in GVA**
   - NO_2 pollution remains **relevant for advanced economies** (in contrast with “night lights”)

3. **Signalling power** of satellite data allows for swifter detection of turning points

4. Satellite data might be valuable for **developing countries** where official statistics are scarce

PAPER AVAILABLE ON SSRN:

HTTPS://PAPERS.SSRN.COM/ABSTRACT_ID=3967146

SCRIPT AVAILABLE ON GITHUB:

HTTPS://GITHUB.COM/THOMASPICAL/SENTINEL5_NO2

APPENDIX

MOTIVATIONS
RELATED LITERATURE AND CONTRIBUTIONS (1/2)

NO_2 pollution

- Economic growth increases NO_2 pollution and conversely economic crisis **lowers NO_2 emissions**: e.g., Boersma and Castellanos (2012) during GFC, Le et *al.* (2020) during COVID-19
- Tracking of **shipping lanes** (Franke *et al.*, 2009) with evidence of a fall during the GFC (de Ruyter de Wilt *et al.*, 2012)

First effort – to the best of our knowledge – in ***using NO_2 pollution to forecast economic variables***

Satellite data

- Use of **"night lights"** to develop alternative measures of GDP (Henderson *et al.*, 2012) or track economic events such as the Covid-19 crisis (Beyer *et al.*, 2021)
- But evidence of a **null elasticity** of economic activity to "night lights" in advanced economies (World Bank, 2017; Hu and Yao, 2019)

Evidence that NO_2 pollution remains a relevant indicator of economic activity for ***advanced countries***

MOTIVATIONS
RELATED LITERATURE AND CONTRIBUTIONS (2/2)

High-frequency data

- Several **alternative high-frequency datasets** emerging during the Covid-19 crisis: e.g., daily credit card spending (Carvalho *et al.*, 2020) or hourly electricity use (Chen *et al.*, 2020)
- NO_2 used as a **high-frequency proxy for economic activity** in some papers such as Deb *et al.* (2020) or Bricongne *et al.* (2020)

Dataset with **global and uniform coverage**, as well as **homogeneous quality** across countries

Forecasting industrial production

- **PMIs widely used** to forecast industrial production (Bruno and Lupi, 2003; Tsuchiya, 2014) including recently in emerging markets (Akdag *et al.*, 2020; Herwadkar and Ghosh, 2020)
- Evidence that PMI-based models **perform better than competing benchmarks** (Bulligan *et al.*, 2010)

Potential of **NO_2 pollution data** while the bulk of the literature has relied on surveys notably PMIs

DATA
STEP 5: INTERPOLATION OF MISSING DATA

Initial issue

- **Large share of missing data** for a locality (up to 50%) with data not missing at random
- If ignored in aggregation, might result in **composition effects** (e.g. if data for an industrial zone are missing)

Literature

- Full-fledged strand in the **geostatistical science** using techniques from linear interpolation (Zhang *et al.*, 2017) to neural networks (Fouladgar and Främling, 2020)
- Three **broad approaches** for interpolation using: external data (such as weather data), spatial correlation ("kriging": Laslett, 1994) and both spatial and temporal correlations (spatiotemporal "kriging": Tadic *et al.*, 2017) – latter found to be the **most performant** (Yang and Hua, 2018)

In the paper

- High **computational cost** from most sophisticated spatiotemporal algorithms (Weiss *et al.*, 2014; Kianan *et al.*, 2021) – not suited for daily and global data
- To balance accuracy and computational cost, implement a **K-Nearest Neighbours algorithms** (KNN) using time and space as the two axis – as in Poloczek *et al.* (2014)
- Hyper-parameters set by 10-fold cross-validation; ***K=26*** minimizes the out-of-sample error

DATA
STEP 6: WEATHER NORMALIZATION

Initial issue

- Air pollution very **sensitive to weather conditions** affecting chemical process of pollutant formation and human polluting behaviours (Rao and Zurbenko, 1994)
- Influence of weather might be **greater than the effect of policies or economic events** (Anh et al., 1997)

Literature

- Again a full-fledged strand in the **geostatistical science** with techniques from OLS regression (Henneman *et al.*, 2015) to gradient boosting techniques (Petetin *et al.*, 2020)
- **Non-linear impact** of weather variables including **interactions** between them (Grange *et al.*, 2018)
- Weather-normalization should be performed in **an homogenous region** given importance of topology (e.g. mountains, coasts) and differences in weather across regions (Liu *et al.*, 2020)

In the paper

- Resort to a **random forest algorithm** to take into account non-linearities and interactions, limit the computational cost, and allow for a low sensitivity to hyper-parameters (Biau and Scornet, 2016) – calibration by out-of-bag process on a number of representative regions
- Run the weather-normalization **by region**

NOWCASTING
MODELLING ISSUES (NO_2 POLLUTION)

Stationarity

- For high-frequency data, year-on-year growth 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 **month-on-month difference of moving averages**

Mixed-frequencies

- Monthly industrial production / PMIs and daily NO_2 pollution

→ **Moving average over one month**

→ Also build on **MIDAS model class** (Ghysels *et al.*, 2004) that allows to

- Put different weight on different lags – first weeks might be more important for month-on-month growth
- Bring lots of lags while preserving parsimony if weighing function independent of the number of lags

Limited time sample

- Data from end-2018 onwards (TROPOMI instrument in action only since this date)

→ Rely on **panel estimates** since data is available for all countries – using panel-MIDAS framework recently introduced by Khalaf *et al.* (2021)

NOWCASTING
HETEROGENEITIES: “CRISIS” *VS.* “NORMAL” PERIODS

Rationale

- High-frequency data might provide useful timely signal during **“crisis” episodes** but might be of second order during **“normal”** periods when conditions remain broadly stable (signal-to-noise ratio)

⇨ Short timespan: instead of breaking by time periods, distinguish by countries depending on their **maximum decline in industrial growth throughout 2020**

	Out-of-sample RMSE relative to the AR(1)=1, by quartile			
	Q1	Q2	Q3	Q4
Pollution-based	**0.58**	**0.55**	**0.80**	**0.98**
PMI-based	0.67	0.60	0.88	1.18

⇨

*Greater accuracy gains for the **most affected countries***

*Pollution-based model **always out-performs benchmarks***

NOWCASTING

HETEROGENEITIES: ADVANCED *VS.* DEVELOPING ECONOMIES

Rationale

- Explanatory power of satellite data can depend on the **level of development** – as in Hu and Yao (2019) for “night lights” not significant for advanced economies

⇨ Interact NO_2 pollution with the **share of manufacturing in the value added**

⇨ Introduce a triple interaction with a **dummy for emerging / developing economies** to test for heterogeneities beyond means of production (e.g. transportation, heating)

	(1)	(2)	(3)
$constant$	0.000 (0.002)	0.000 (0.002)	0.000 (0.002)
$dlog(\sum_{j=1}^{31} Pol_{i,t-j})$	0.309*** (0.021)	0.431*** (0.069)	0.502*** (0.125)
$dlog(\sum_{j=1}^{31} Pol_{i,t-j}) \cdot manuf_i$		-0.770* (0.439)	-2.048** (0.920)
$dlog(\sum_{j=1}^{31} Pol_{i,t-j}) \cdot \delta_i^{EME}$			0.164 (0.157)
$dlog(\sum_{j=1}^{31} Pol_{i,t-j}) \cdot manuf_i \cdot \delta_i^{EME}$			0.517 (1.064)
Country fixed effects	Yes	Yes	Yes
Adjusted R^2	0.25	0.26	0.30

⇨

Share of manufacturing significant: the higher the share, the higher the elasticity

*Dummy for emerging and developing economies **not significant***

CONCLUSION
NEXT STEPS

1. Confirm the value-added of NO_2 pollution data during **“normal” episodes** – more generally confirm results over a longer time period given peculiarities of the Covid-19 crisis

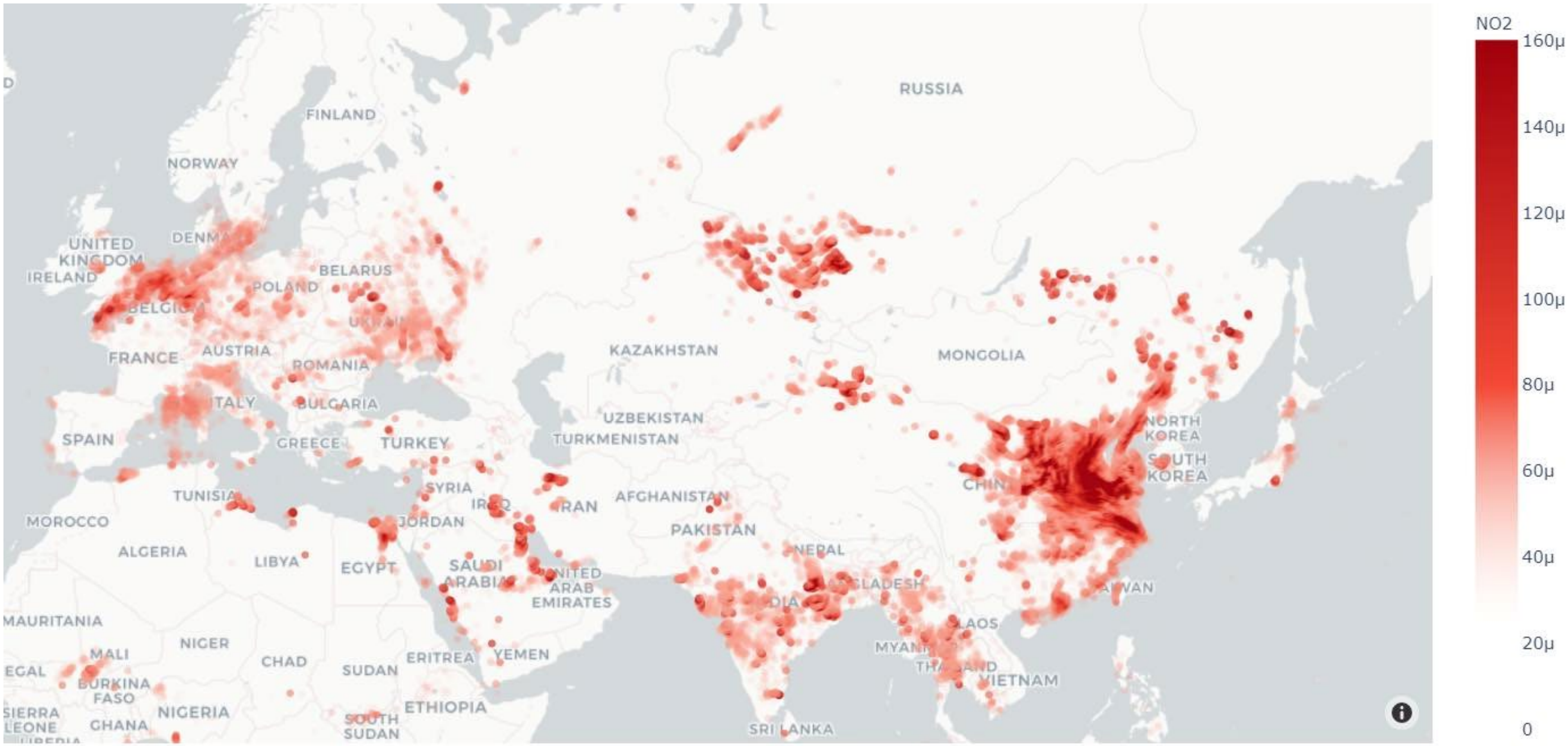
2. Exploit the granularity of satellite data to derive indices of economic activity at **local level**

3. Explore the potential of pollution data to track **other macroeconomic variables** (e.g. trade)

4. Satellite data might be used to develop a PMI-like indicator in **developing countries** not covered by PMI surveys – possibly combined with other alternative (e.g. Google Trends)

5. Other types of satellite data might be exploited – e.g. **infrared data** to detect the heat produced by factories (on-going work)

VISUALISATION OF RAW DATA



BANQUE DE FRANCE
EUROSYSTÈME

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

NON LINEARITIES IN WEATHER CORRECTION

Partial dependencies plots for meteorological variables

Sources: ESA, WAQI, NOAA, authors' calculations