

Chan, Ho Fai; Skali, Ahmed; Torgler, Benno

**Working Paper**

## A global dataset of human mobility

CREMA Working Paper, No. 2020-04

**Provided in Cooperation with:**

CREMA - Center for Research in Economics, Management and the Arts, Zürich

*Suggested Citation:* Chan, Ho Fai; Skali, Ahmed; Torgler, Benno (2020) : A global dataset of human mobility, CREMA Working Paper, No. 2020-04, Center for Research in Economics, Management and the Arts (CREMA), Zürich

This Version is available at:

<https://hdl.handle.net/10419/225546>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*



Center for Research in Economics, Management and the Arts

# A Global Dataset of Human Mobility

Working Paper No. 2020-04

CREMA Südstrasse 11 CH - 8008 Zürich [www.crema-research.ch](http://www.crema-research.ch)

# Title

## A Global Dataset of Human Mobility

### Authors

Ho Fai Chan<sup>1,2</sup>, Ahmed Skali<sup>3</sup>, Benno Torgler<sup>1,2,4</sup>

### Affiliations

1. School of Economics and Finance, Queensland University of Technology
  2. Centre for Behavioural Economics, Society and Technology (BEST)
  3. Department of Economics, Deakin University
  4. CREMA – Center for Research in Economics, Management and the Arts, Gellertstrasse 18, CH-4052 Basel, Switzerland
- corresponding author(s): Ho Fai Chan (hofai.chan@qut.edu.au)

### Abstract

In the midst of the COVID-19 pandemic currently affecting every corner of the globe, there is a critical need for understanding and mapping human movement in order to formulate appropriate scientific and policy responses. To this end, we provide ready-to-use indexes of human mobility spanning 131 countries and territories and 818 sub-national regions across a time period between 16 February 2020 to 29 March 2020. We extract these data from Google’s newly released mobility reports, which cover phone-tracking-based changes in mobility across several types of locations, including retail and recreation, grocery stores and pharmacies, parks, transit stations, workplaces, and private residences. We make these data immediately available to the scientific community and the broader public at <https://osf.io/rzd8k/> in order to support research into the COVID-19 pandemic.

### Background & Summary

On 2 April 2020, Google released Community Mobility Reports, outlining changes in mobility for several geographies across types of locations, including retail and recreation, grocery stores and pharmacies, parks, transit stations, workplaces, and private residences. The reports are based on aggregated information from users who have actively turned on their location history setting as the setting is off by default<sup>1</sup>. The reports were released as PDF files, with each file containing several graphs displaying the change in mobility between 16 February 2020 (near the start of the pandemic in most countries in the world) and 29 March 2020. Our motivation for extracting the data from these graphs is two-fold. First, Google has only released graphs embedded in PDF files, rather than raw data. The former are only of limited value to researchers wishing to understand the current pandemic. Second, the PDF files released by Google “will be available for a limited time, so long as public health officials find them useful in their work to stop the spread of COVID-19”<sup>2</sup>. In order to ensure the data are available for all to use in the spirit of open science, we proceeded to extract and convert the data into a ready-to-use format as explained in the following section. As Jen Fitzpatrick (SVP, Geo) and Karen DeSalvo (Chief Health Officer, Google Health) point out, such data could help making critical decisions to combat COVID-19<sup>3</sup>.

### Methods

Each community mobility report was published as a PDF file, with several graphs embedded in each file. Each graph plots the time series of changes in mobility relative to a baseline. The

baseline is the median value for the corresponding day of the week, during the period between 3 January 2020 to 6 February 2020. For example, if the mobility change at parks in New York City is -10 on Monday 23 March, this means that New York City residents were 10 percentage points less likely to find themselves in parks on Monday 23 March than they were on the Mondays between 3 January and 6 February 2020. Absolute number of visits were not released by Google. The aggregated information allows helps to protect people's privacy.

We extract the data from each graph using a combination of WebPlotDigitizer and AutoHotKey. WebPlotDigitizer<sup>4</sup> is an open source tool designed to obtain numerical data from graphs. First, the user manually calibrates the X and Y axes at known values, allowing WebPlotDigitizer to extract the graduation of each axis. Second, the user specifies the color of the curve for which numerical data is of interest. WebPlotDigitizer's auto-detection algorithm then produces numerical data from the curve of interest and the location of each point on the curve with respect to the user-calibrated X-Y axes.

Since we are dealing with a large number of graphs (5,766 in total, with 131 countries and 830 regional level graphs on six location categories), manually calibrating each plot extraction would have been prohibitively expensive. Repetitive manual calibration is also likely to introduce measurement errors, thus lowering the precision and consistency of the extracted data. We automated this task using AutoHotKey, another open source software, allowing for pixel-by-pixel precision control on mouse movement, minimising measurement errors when aligning the (2D) X-Y Axes on WebPlotDigitizer. Regional-level information is also important due to the urgency and regional differences in the coronavirus pandemic. It also allows to better identify and evaluate policy strategies.

We split the pages for each report in order to process images in WebPlotDigitizer. The layout is similar across all reports. The first page consists of graphs for changes in national mobility for *Retail & recreation*, *Grocery & pharmacy*, and *Parks*, and the second page consists of graphs for changes in mobility in *Transit stations*, *Workplaces*, and *Residential*. For 36 countries, regional-level graphs are available from page three of the report onward. Three AutoHotKey scripts were written to process the data extraction for each part of the report. Each script loads the image into WebPlotDigitizer, aligns the X-Y axis, and exports the data. We utilize the default averaging window algorithm<sup>4</sup> as the algorithm for automatic extraction of data points from each graph, with input parameter  $\Delta X$  and  $\Delta Y$  set as 1 pixel (Px) and RGB values (65 133 244) for the color of the line in each graph, which is identical across all reports released. Data points for each graph are then exported into a CSV file and collated using Stata to tag each time series with the correct country/region and location category.

With the fine grid of 1x1 pixel data point detection, the resulting time series captures multiple values around each calendar date (on average there are ~30 points around an integer date). These points are then aggregated into a single value using weighted averaging, assigning higher weights to points with smaller deviations (absolute difference) in x-axis to the date (rounded in integer). Specifically, weights are calculated as the cube of the absolute difference.

We provide a visualization of the data in Figure 1, where we show the changes in mobility for each category over time across the globe. Figure 1 provides a sense of which countries were "early movers" in reducing mobility.

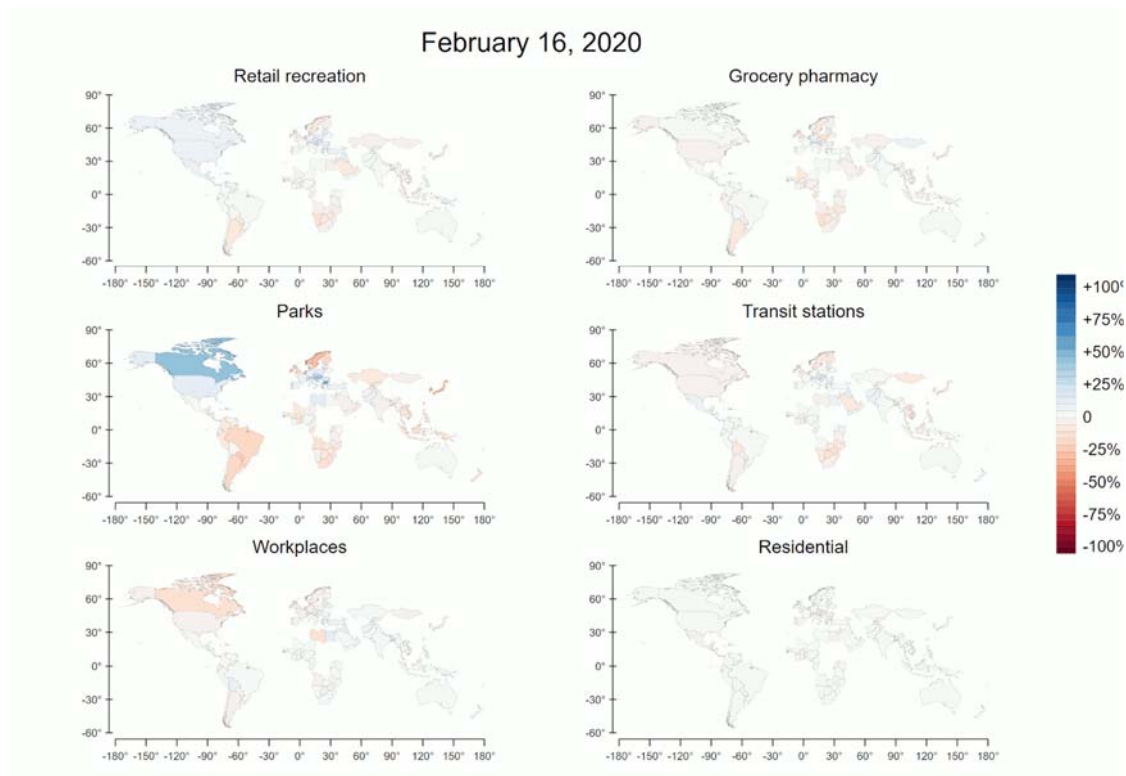


Figure 1. Global visualization of changes in mobility around the world.

## Data Records

The data are available on the Open Science Framework at <https://osf.io/rzd8k/>. Two files are provided in CSV format: *MobilityCountries* is the country-level daily dataset ( $N_{\text{countries}} = 131$ ,  $t = 43$ ,  $N_{\text{total}} = 5,633$ ) and *MobilityRegions* is the regional-level daily dataset ( $N_{\text{regions}} = 830$ ,  $t \leq 43$ ,  $N_{\text{total}} = 35,276$ ). Mobility data are provided by Google for sub-national regions within 36 countries. In the regional dataset, data are not available for all days in 48 of the 830 regions; however, there is full coverage across 94.2% of regions. The variables provided are:

**iso2c:** country identifiers following the ISO 3166-1 alpha-2 convention.

**region:** sub-national region name (regional dataset only).

**date:** calendar date.

**Retail\_recreation; Grocery\_pharmacy; Parks; Transit\_stations; Workplaces; Residential:** mobility indexes for the relevant location type as described in Methods section above.

**not\_enough\_data:** indicator for time series with “Not enough data for this date” tagged in the Mobility report.

## Technical Validation

While prior research has demonstrated the high validity of using WebPlotDigitizer for data extraction (e.g., high levels of precision, intercoder reliability, and validity<sup>5</sup>), we examine the accuracy of our results by implementing our data extraction procedure with a known set of similarly formatted data. Specifically, we obtained graphs from Google Trends (<https://trends.google.com/trends/>), which generates graphs in a graphical style similar to the

Mobility reports. We performed searches on Google Trends for five broad topics and exported the resulting graphs into PDFs along with the actual data values from Google Trends (see GM\_Validation section in OSF link). We then performed the same data extraction procedure using WebPlotDigitizer and AutoHotKey and compared the extracted data series with the actual data series. The results show that the correlation between the two series have a very strong positive correlation (average Pearson's  $\rho = 0.987$ ; smallest  $\rho = 0.983$ ). The average percentage deviation from the actual data series is about 2.5% (SD = 3.09).

## Code Availability

All code used in this study is publicly available on OSF at <https://osf.io/rzd8k/>.

WebPlotDigitizer: Version 4.2, released 7 April 2019.

AutoHotKey: Version v1.1.22.07

Stata16MP: Version 16.1

## Author contributions

**Ho Fai Chan:** Methodology, Software, Validation, Formal analysis, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration.

**Ahmed Skali:** Conceptualization, Methodology, Validation, Resources, Writing - Original Draft, Writing - Review & Editing, Project administration. **Benno Torgler:** Conceptualization, Methodology, Investigation, Writing - Original Draft, Writing - Review & Editing, Project administration.

## Competing interests

We have no conflicting / competing interests.

## References

1. Google, Manage your Location History, <https://support.google.com/accounts/answer/3118687> (2020)
2. Google, COVID-19 Mobility Reports, <https://www.google.com/covid19/mobility/>
3. Fitzpatrick, J. & DeSalvo, K. Helping public health officials combat COVID-19, <https://www.blog.google/technology/health/covid-19-community-mobility-reports> (2020)
4. Rohatgi, A. WebPlotDigitizer, <https://automeris.io/WebPlotDigitizer/citation.html> (2019)
5. Drevon, D., Fursa, S. & Malcolm, A. Intercoder Reliability and Validity of WebPlotDigitizer in Extracting Graphed Data. *Behav. Modif.* **41**, 323-339 (2017)