GTFS2EMIS an R package to estimate public transport emissions based on GTFS data

Rafael H.M. Pereira[1, *], Pedro R. Andrade[2], and Joao Pedro Bazzo[1, *]

[1]Institute for Applied Economic Research, Brasília, Brazil
[2]National Institute of Space Research, São José dos Campos, Brazil

Correspondence: Joao Pedro Bazzo (joao.bazzo@gmail.com)

Abstract. The abstract goes here. It can also be on *multiple lines*.

1 Introduction

Transport emissions have been widely recognized among the leading and growing contributors to global emissions (Caiazzo et al., 2013; Nocera et al., 2018,). There is also scant evidence of how transportation activities impact air quality in cities (Landrigan et al., 2018), aggravating the negative short- and long-term effects pollution has on children premature deaths (Currie and Walker, 2009), cardiovascular diseases (Brook et al., 2010; Turner et al., 2016), ischemic stroke (Wellenius et al., 2012) and cognitive development (Chen et al., 2017; Fu et al., 2019; Shehab and Pope, 2019; Zhang et al., 2018). This has a higher impact in low and middle-income countries (Combes and Franchineau, 2019), especially in more vulnerable groups, such as elderly people (Yap et al., 2019) and children (Braga et al., 2001; Gauderman et al., 2015). Understanding the spatial and temporal patterns of emissions is a key factor to adopt more precise public policies to reduce air pollution exposure (Clark et al., 2014; Targino et al., 2018), improve the housing and land use planning, and subsidize public investments Particularly with public transport policies, extensive efforts have been made to quantify air pollution patterns and to understand the impact of fleet technology, fuel, topography, and route choices on overall levels of air pollution (refs).

Most of the studies on public transportation emissions focus on emissions monitoring with remote sensing devices (refs), onboard diagnosis using PEMS (refs), vehicle counts data (refs), radar equipments (refs), public transport GPS data (refs), and fuel based estimates (top down approaches). However, due to a restriction of spatial application, or costly local data collection, some of these approaches might not be easy to scale for other cities. Other approaches, such as air quality analysis (refs), satellite data observations (refs) would represent the final state of air quality but ultimately do not look only at bus emissions. Studies that identify and map pollution in low and middle-income countries are still needed, particularly when easily reproduced.

Our analysis contributes to the public transport emission literature by developing a novel methodology that leverages a widespread open data format for public transport networks known as General Transit Feed Specification (GTFS) to estimate vehicle emissions in a high spatial and temporal resolution. The method is presented in a R package "gtfs2emis" to allow friendly and reproducible use for different applications. In this paper, we present estimates on CO2, NOx, PM10, CO, CH4 hot exhaust emissions for 24 cities in the world, considering a typical business day on October 2019 — as a baseline date. A comparison of the vehicle emissions rates of those transport networks is also presented, along with a discussion on the versatility and limitations of the method.

The remainder of this paper is organized as follows. Section two presents a brief literature review on public transportation emissions. Section three describes the data and methods for the public transport data, emission factors databases and vehicle emissions estimates.. Section four presents the results, looking at the overall levels of pollutant emissions and its spatial and temporal resolution behavior. Finally, section five presents the final remarks and discusses some environmental insights that can be drawn from the package.

2 Methods

2.1 Transport model

2.1.1 GTFS data

The method used in the paper leverages a global standard format for public transport data known as General Transit Feed Specification (GTFS). The GTFS data of a given public transport system is collected in a ZIP file, that brings detailed geolocated information on scheduled services including its stops, routes, trips, stop times and calendar organized in a structured text file.

This format was originally created in the mid-2000s by Google and Portland transport authority and it has since then become adopted worldwide by more than 600 cities (noa, b). The simplicity of this data format and its wide adoption creates a common ground for researchers and practitioners across the globe to develop and share computational methods that can be seamlessly deployed across multiple cities to manage fleet allocation, plan transport services, etc. Detailed aspects of the GTFS contents are found in (noa, c).

We expand the sample of cities with GTFS data collected directly from local transport authorities and from Transit Feed archives - the largest GTFS open data repository website. The GTFS data were pre-processed to filter by bus routes and to check consistencies on the required files (e.g. agency, stops, routes, trips, stop times). The months of October and November 2019 are used as the baseline date for the emissions estimates, in order to avoid possible differences in public transportation schedules due to the COVID-19 pandemic. A summary of the main GTFS information is described in Table 1, together with the fleet data.

2.2 GTFS2gps

The process to convert GTFS files to GPS-like spatial data points is done for every trip of the schedule. It interpolates the space-time position of each vehicle in each trip considering the network distance and average speed between stops. It samples the timestamp of each vehicle every 50m.

Once a GPS data format is generated from GTFS, a few adjustments on speed is made. Since the mean speed is estimated in a street link between two bus stops (e.g. stop "i" and "i+1"), the data on departure time of "i" and arrival time of "i+1" is necessary. If "i+1" arrival time is very close or the same as the departure time of "i", the speed will be very high (>100 kph) or infinite. In such cases, we consider the speed to be the mean speed of all the trips in this specific bus route. Therefore, the departure times of the GPS are recalculated based on the new travelled speeds. This adjustment is also necessary in street links without valid arrival or departure times: before the first bus stop of the route (where only arrival time is available); after the last bus stop (only departure time is available); and trips with invalid stop times.

2.3 Emission model

2.3.1 Emission factors

Emission factors (EFs) are empirical functional relations between pollutant emissions and the activity that causes them (Franco et al., 2013). This study uses EF based on Environmental Sanitation Technology Company (CETESB) — for Brazil; California Air Resources Board (CARB) — for the California U.S; Motor Vehicle Emission Model (MOVES) for regions in U.S. outside California; European Environment Agency (EEA) — for European countries. Among these sources, EMFAC, MOVES and EMEP/EEA have speed-dependent emission factors. For the Brazilian cities' estimates, the local EF is scaled to the EMEP/EEA function considering the average speed of the vehicle driving cycle. The brazilian emission factor is related to speed by the expression

$$EF^{scaled}_{i,j,k,l} = EF^{local}_{i,j,k,l}\frac{EF(V_i)_{j,k,l}}{EF(V_{dc})_{j,k,l}}. \tag{1}$$

where $EF^{scaled}_{i,j,k,l}$ is the scaled emission factors for each street link i, bus type j, fuel k, age l,; $EF^{local}_{i,j,k,l}$ is the local emission factor; $EF(V_i)_{j,k,m}$ and $EF(V_{dc})_{j,k,l}$ is the EEA emission factor at the speed of V_i;and the driving cycle V_{dc} of FTP-25 (33 kph).

The CARB's (California Air Resources Board) EMFAC (EMission FACtor) model estimates statewide and regional emissions inventory by multiplying emissions rates with vehicle activity data. The running exhaust emission factors are distributed according to speed bins from 5 to 90, with 5 mph incremental. It allows modeling one season (summer, winter or annual average) due to temporal variations of EF due to meteorological conditions (temperature and relative humidity); and geographic area (Statewide, Air Basin, Air District, Metropolitan Planning Organization, Country and Sub-Area). Detailed data on emission factors are found in noa (a).

Table 1. Emission factor sources, related bus categories and variables associated.

Source	Buses categories	Variables
CETESB	Micro, Standard, Articulated	Age, Fuel, EURO stage
EMFAC model	Urban Buses	Age, Fuel
EMEP/EEA	Micro, Standard, Articulated	Fuel, EURO stage, technology, load, slope
MOVES U.S EPA	Urban Buses	Age, Fuel

Table footnotes

Table 2. Association between model year, PROCONVE Norm and EURO stage for the brazilian fleet (adapted from ICCT).

PROCONVE Norm	EURO equivalent	Year range
P4	II	1998 - 2003
P5	III	2004 - 2011
P8	V	2012 - 2019

Table footnotes

The Motor Vehicle Emission Simular (MOVES) is a set of tools for modelling air pollution emissions generated by onroad and nonroad mobiles sources. It has been the official model for transportation conformity analysis outside California state. The MOVES database of emission rates are stored according to pollutant, emission process (e.g. running exhaust, start exhaust, extended idle), fuel type, regulatory class, model year, operation mode and vehicle age. Urban buses are classified into fuel type (CNG, Diesel, Gasoline), age, speed range, pollutant, and source type. Detailed discussion on heavy-duty emission rates are found in [EPA-HDV-MANUAL].

European exhaust emission factors are presented in the EMEP/EEA air pollutant inventory guidebook, considering the COPERT 5.4 Software version. This study considered average load on all trips and zero slope rate for all studied cities. Urban Buses are presented in three main categories, which are Urban Buses Midi (<= 15 t), Standard (15 - 18 t), Articulated (>18 t), Urban Diesel Buses Hybrid, CNG (Compressed Natural Gas) Buses. The speed dependent emissions factors for diesel urban buses have been taken from HBEFA (Handbook Emission Factors for Road Transport), for Euro I to Euro VI emissions standards. Distinct parameters of EF were considered for Euro V standard, according to the control technology, that can be Exhaust Gas Recirculation — EGR or Selective Catalytic Reduction — SCR. According to EMEP/EEA, it is estimated that 75% of Euro V heavy-duty vehicles are equipped with SCR. For the category of CNG buses, it has an additional emission standard known as Enhanced Environmental Vehicles (EEV), since it may have different combustion, after-treatment technology, and are associated with lower PM and NOx emission rates compared with diesel buses (add source). Only older CNG buses are classified in EURO I, II, or III. Detailed aspects of hot exhaust emissions rates of Europe are found in the EMEP/EEA guidebook.

Table 3. Summary of all vehicle classes covered by the Tier 2 methodology

Urban bus category	Euro Stage
Urban CNG buses	Euro I, Euro II, Euro III, EEV
Urban buses	Conventional, Euro I - Euro VI
Urban Diesel Hybrid	Euro VI
Urban biodiesel buses	Conventional, Euro I - Euro VI

Table footnotes

2.3.2 Emission estimates

The estimates of hot exhaust emissions for each trip is given by

$$EH_{i,j,k,l} = L_i \times EF_{i,j,k,l} \tag{2}$$

where $EH_{i,j,k,l}$ is the emission for the street link i, vehicle category j, fuel k and age l; L_i is the length of the street link i(km); $EF_{i,j,k,l}$ is the emission factor (g/km). In order to evaluate emissions spatially, the overall emissions are allocated into a 350m resolution grid (H3 resolution from Uber).

2.3.3 Emission post-processing

Post-processing emissions are auxiliary functions used to analyze the spatial and temporal patterns. These function are:

1. `emis_summary`: Aggregate emissions generated by the emis function by time, vehicle type and segment link (spatial).
2. `emi_to_dt`: Convert emissions estimates from list to data.table format
3. `emis_grid`: Aggregate emissions proportionally into grid cells through an intersection operation

3 São Paulo urban bus emissions using GTFS2emis model

3.1 Traffic data

GTFS data of São Paulo city were retrivied in SPTRAN (São Paulo Transportation Agency) and EMTU (Metropolitan Transportation Agency). SPTRAN database includes all trips inside São Paulo territory while EMTU database presents intercity level buses routes, where some of them runs partially within SP. Considering the metropolitan scale of São Paulo city, we considered all buses routes that crosses or is within the city boundary.

The merged GTFS were filtered by route_id and service's_id, in order to estimate only buses trips in only day (monday).

```
library(data.table)
```

```
library(magrittr)
gtfs_path <- "../../../Dropbox/IPEA/gtfs2gps/data-raw/gtfs/gtfs_spo_sptrans_2019-10.zip"

tmp_gtfs <- gtfs2gps::read_gtfs(gtfs_path)
tmp_gtfs <- gtfs2gps::filter_by_day(gtfs_data = tmp_gtfs, days = "monday")

# filter by bus route
temp_routeid <- tmp_gtfs$routes[route_type == 3,route_id]
temp_shapeids <- tmp_gtfs$trips[route_id %in% unique(temp_routeid),shape_id]
tmp_gtfs <- gtfs2gps::filter_by_shape_id(gtfs_data = tmp_gtfs,
                                         shape_ids = unique(temp_shapeids))
```

Viewing GTFS file

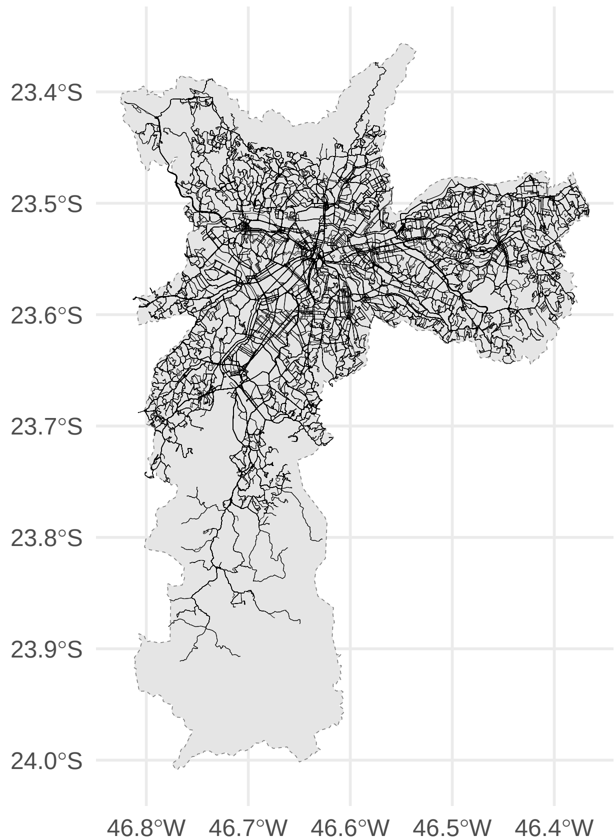
```
library(ggplot2)
library(sfheaders)
library(sf)
library(magrittr)

boundary <- geobr::read_municipality(code_muni = 3550308)
```

```
## Downloading: 770 B     Downloading: 770 B     Downloading: 1.6 kB     Downloading: 1.6 kB
```

```
tmp_shp <- sfheaders::sf_multilinestring(obj = tmp_gtfs$shapes,
                                         x = "shape_pt_lon",
                                         y = "shape_pt_lat",
                                         multilinestring_id = "shape_id") %>%
  sf::st_set_crs(4326)


ggplot() +
  geom_sf(data = boundary,color = "grey50",lty = 2,size = 0.15) +
  geom_sf(data = tmp_shp,size = .02) +
  theme_minimal()+
  theme(legend.position = "none")+
  labs(color = "Boundary")
```




```
#mapview::mapview(tmp_shp, zcol = "shape_id") +
#  mapview::mapview(boundary$geom)
```

3.2 Fleet data

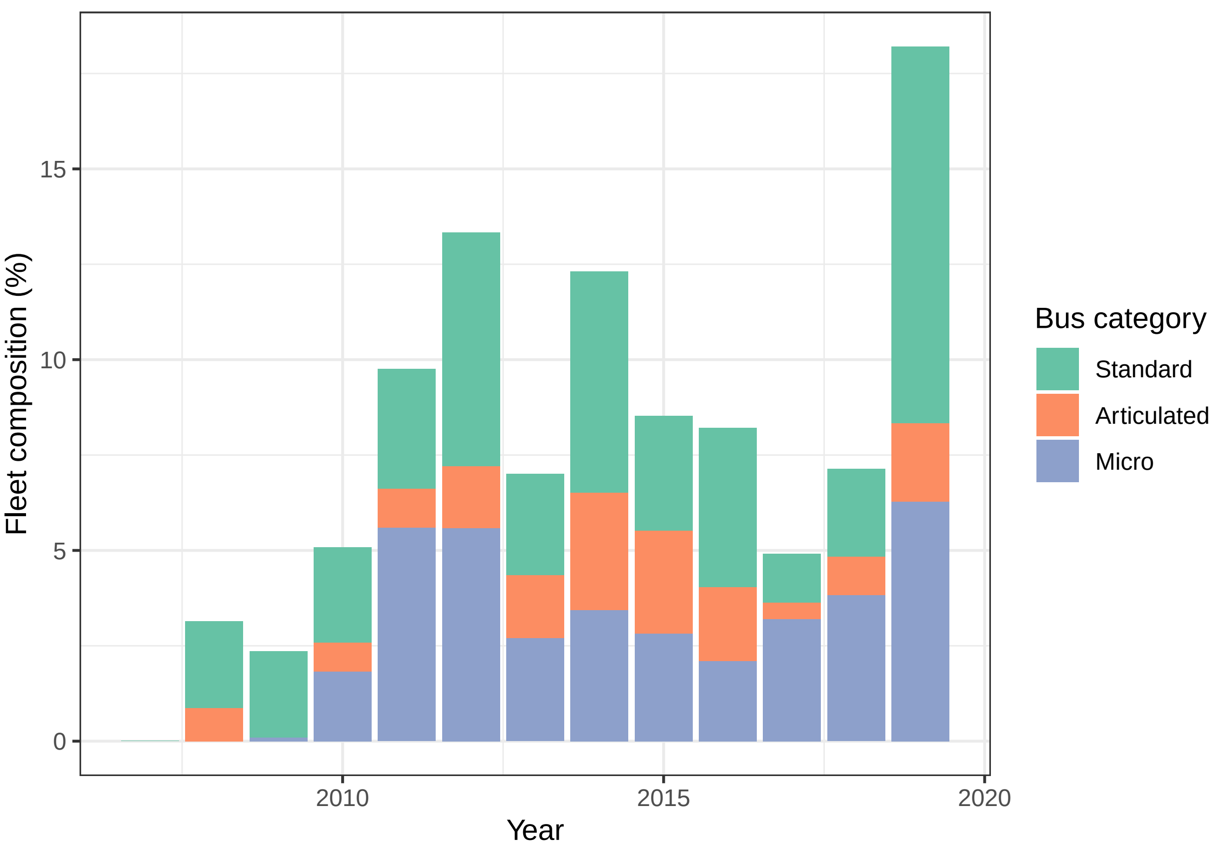
The data on fleet were obtained in SPTRANS and EMTU database. The relevant information used to compute emissions were age, fuel, category, and route in which buses uses to run. This last data were not available for SP. The approach used to allocate buses per route was considering the all the routes have the same distribution of vehicles, which is based on fleet type frequency.

The figure below shows the fleet of São Paulo, according to fuel, age and category.

```
library(ggplot2)
library(magrittr)
library(data.table)
tmp_fleet <- readr::read_rds("../../../Dropbox/IPEA/gtfs2gps/data-raw/fleet/spo/bra_spo.rds"

tmp_fleet %>%
  .[,lapply(.SD,sum),.SDcols = "fleet_composition",by = .(year,type_name_br)] %>%
```

```
.[order(year),] %>%
.[,type_name_br_lab := factor(x = type_name_br
                              ,levels = c("BUS_URBAN_D",
                                          "BUS_ARTIC_D",
                                          "BUS_MICRO_D")
                              ,labels = c("Standard",
                                          "Articulated",
                                          "Micro"))] %>%
ggplot() +
geom_bar(aes(x = year,y = 100 * fleet_composition,fill = type_name_br_lab)
         , stat = "identity")+
scale_fill_brewer(palette = "Set2") +
labs(y = "Fleet composition (%)"
     , x = "Year"
     , fill = "Bus category") +
theme_bw()
```



3.3 Emission estimates

Figure 1 shows the emissions of PM10 in the urban area of São Paulo, with a spatial resolution of 350 m (Uber Hexagon 09). The higher values of emissions indicates streets and intersections of high Vehicle Traveled Kilometers (VTK). Intersections with high VTK are usually interpreted as streets intersections or buses terminals.

Figure 2 shows the PM10 emissions distributed hourly. Higher values of emissions on morning and afternoon peak indicates higher frequency of buses. It can be also associated with average lower speeds.

4 Results and discussion

5 Conclusions

6 Discussions and Conclusions

Given the wide use of GTFS to perform transportation analysis worldwide, the package `gtfs2emis` can promote easier estimates of tailpipe emissions. Detailed input data can improve the quality of estimates.

Code and data availability. use this to add a statement when having data sets and software code available

Sample availability. use this section when having geoscientific samples available

Video supplement. use this section when having video supplements available

Appendix A: Figures and tables in appendices

Regarding figures and tables in appendices, the following two options are possible depending on your general handling of figures and tables in the manuscript environment:

A1 Option 1

If you sorted all figures and tables into the sections of the text, please also sort the appendix figures and appendix tables into the respective appendix sections. They will be correctly named automatically.

A2 Option 2

If you put all figures after the reference list, please insert appendix tables and figures after the normal tables and figures.

To rename them correctly to A1, A2, etc., please add the following commands in front of them: `\appendixfigures` needs to be added in front of appendix figures `\appendixtables` needs to be added in front of appendix tables

Please add `\clearpage` between each table and/or figure. Further guidelines on figures and tables can be found below.

Author contributions. Daniel wrote the package. Josiah thought about poterry. Markus filled in for a second author.

Competing interests. The authors declare no competing interests.

Disclaimer. We like Copernicus.

Acknowledgements. Thanks to the rticles contributors!

References

EMFAC, https://arb.ca.gov/emfac/project-analysis, a.

OpenMobilityData - Public transit feeds from around the world, https://transitfeeds.com/, b.

Reference | Static Transit | Google Developers, https://developers.google.com/transit/gtfs/reference, c.

Braga, A. L., Saldiva, P. H., Pereira, L. A., Menezes, J. J., Conceição, G. M., Lin, C. A., Zanobetti, A., Schwartz, J., and Dockery, D. W.: Health effects of air pollution exposure on children and adolescents in São Paulo, Brazil, Pediatric Pulmonology, 31, 106–113, https://doi.org/10.1002/1099-0496(200102)31:2<106::aid-ppul1017>3.0.co;2-m, 2001.

Brook, R. D., Rajagopalan, S., Pope, C. A., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., Holguin, F., Hong, Y., Luepker, R. V., Mittleman, M. A., Peters, A., Siscovick, D., Smith, S. C., Whitsel, L., and Kaufman, J. D.: Particulate Matter Air Pollution and Cardiovascular Disease, Circulation, 121, 2331–2378, https://doi.org/10.1161/CIR.0b013e3181dbece1, publisher: American Heart Association, 2010.

Caiazzo, F., Ashok, A., Waitz, I. A., Yim, S. H. L., and Barrett, S. R. H.: Air pollution and early deaths in the United States. Part I: Quantifying the impact of major sectors in 2005, Atmospheric Environment, 79, 198–208, https://doi.org/10.1016/j.atmosenv.2013.05.081, 2013.

Chen, H., Kwong, J. C., Copes, R., Tu, K., Villeneuve, P. J., Donkelaar, A. v., Hystad, P., Martin, R. V., Murray, B. J., Jessiman, B., Wilton, A. S., Kopp, A., and Burnett, R. T.: Living near major roads and the incidence of dementia, Parkinson's disease, and multiple sclerosis: a population-based cohort study, The Lancet, 389, 718–726, https://doi.org/10.1016/S0140-6736(16)32399-6, publisher: Elsevier, 2017.

Clark, L. P., Millet, D. B., and Marshall, J. D.: National Patterns in Environmental Injustice and Inequality: Outdoor NO2 Air Pollution in the United States, PLOS ONE, 9, e94 431, https://doi.org/10.1371/journal.pone.0094431, publisher: Public Library of Science, 2014.

Combes, A. and Franchineau, G.: Fine particle environmental pollution and cardiovascular diseases, Metabolism: Clinical and Experimental, 100S, 153 944, https://doi.org/10.1016/j.metabol.2019.07.008, 2019.

Currie, J. and Walker, W. R.: Traffic Congestion and Infant Health: Evidence from E-ZPass, Tech. Rep. w15413, National Bureau of Economic Research, https://doi.org/10.3386/w15413, 2009.

Fu, P., Guo, X., Cheung, F. M. H., and Yung, K. K. L.: The association between PM2.5 exposure and neurological disorders: A systematic review and meta-analysis, The Science of the Total Environment, 655, 1240–1248, https://doi.org/10.1016/j.scitotenv.2018.11.218, 2019.

Gauderman, W. J., Urman, R., Avol, E., Berhane, K., McConnell, R., Rappaport, E., Chang, R., Lurmann, F., and Gilliland, F.: Association of Improved Air Quality with Lung Development in Children, New England Journal of Medicine, 372, 905–913, https://doi.org/10.1056/NEJMoa1414123, publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMoa1414123, 2015.

Landrigan, P. J., Fuller, R., Acosta, N. J. R., Adeyi, O., Arnold, R., Basu, N. N., Baldé, A. B., Bertollini, R., Bose-O'Reilly, S., Boufford, J. I., Breysse, P. N., Chiles, T., Mahidol, C., Coll-Seck, A. M., Cropper, M. L., Fobil, J., Fuster, V., Greenstone, M., Haines, A., Hanrahan, D., Hunter, D., Khare, M., Krupnick, A., Lanphear, B., Lohani, B., Martin, K., Mathiasen, K. V., McTeer, M. A., Murray, C. J. L., Ndahimananjara, J. D., Perera, F., Potočnik, J., Preker, A. S., Ramesh, J., Rockström, J., Salinas, C., Samson, L. D., Sandilya, K., Sly, P. D., Smith, K. R., Steiner, A., Stewart, R. B., Suk, W. A., Schayck, O. C. P. v., Yadama, G. N., Yumkella, K., and Zhong, M.: The Lancet Commission on pollution and health, The Lancet, 391, 462–512, https://doi.org/10.1016/S0140-6736(17)32345-0, publisher: Elsevier, 2018.

Nocera, S., Ruiz-Alarcón-Quintero, C., and Cavallaro, F.: Assessing carbon emissions from road transport through traffic flow estimators, Transportation Research Part C: Emerging Technologies, 95, 125–148, https://doi.org/10.1016/j.trc.2018.07.020, 2018.

Shehab, M. A. and Pope, F. D.: Effects of short-term exposure to particulate matter air pollution on cognitive performance, Scientific Reports, 9, 8237, https://doi.org/10.1038/s41598-019-44561-0, number: 1 Publisher: Nature Publishing Group, 2019.

Targino, A. C., Krecl, P., Danziger Filho, J. E., Segura, J. F., and Gibson, M. D.: Spatial variability of on-bicycle black carbon concentrations in the megacity of São Paulo: A pilot study, Environmental Pollution (Barking, Essex: 1987), 242, 539–543, https://doi.org/10.1016/j.envpol.2018.07.003, 2018.

Turner, M. C., Jerrett, M., Pope, C. A., Krewski, D., Gapstur, S. M., Diver, W. R., Beckerman, B. S., Marshall, J. D., Su, J., Crouse, D. L., and Burnett, R. T.: Long-Term Ozone Exposure and Mortality in a Large Prospective Study, American Journal of Respiratory and Critical Care Medicine, 193, 1134–1142, https://doi.org/10.1164/rccm.201508-1633OC, publisher: American Thoracic Society - AJRCCM, 2016.

Wellenius, G. A., Burger, M. R., Coull, B. A., Schwartz, J., Suh, H. H., Koutrakis, P., Schlaug, G., Gold, D. R., and Mittleman, M. A.: Ambient Air Pollution and the Risk of Acute Ischemic Stroke, Archives of Internal Medicine, 172, 229–234, https://doi.org/10.1001/archinternmed.2011.732, 2012.

Yap, J., Ng, Y., Yeo, K. K., Sahlén, A., Lam, C. S. P., Lee, V., and Ma, S.: Particulate air pollution on cardiovascular mortality in the tropics: impact on the elderly, Environmental Health, 18, 34, https://doi.org/10.1186/s12940-019-0476-4, 2019.

Zhang, X., Chen, X., and Zhang, X.: The impact of exposure to air pollution on cognitive performance, Proceedings of the National Academy of Sciences, 115, 9193–9197, https://doi.org/10.1073/pnas.1809474115, publisher: National Academy of Sciences Section: Social Sciences, 2018.