ENABLING AIR QUALITY MONITORING WITH THE OPEN DATA CUBE: IMPLEMENTATION FOR SENTINEL-5P AND GROUND SENSOR OBSERVATIONS

J. R. Cedeno Jimenez*[1], D. Oxoli[1], M. A. Brovelli[1]

[1] Department of Civil and Environmental Engineering, Politecnico di Milano, P.zza Leonardo da Vinci 32, 20133, Milan, Italy - jesusrodrigo.cedeno@mail.polimi.it, (daniele.oxoli, maria.brovelli)@polimi.it

Commission IV

KEY WORDS: Air Pollution, Air Quality Sensors, Sentinel-5P, Earth Observations, Open Data Cube, Nitrogen Dioxide

ABSTRACT:

Nowadays, the amount of open geospatial data delivered e.g. by private and public entities, such as environmental agencies, enables outstanding possibilities to any user interested in investigating real-world phenomena. However, the availability of such information presents several challenges in terms of its practical use. These are mainly connected to the heterogeneity of data sources, formats and processing tools which have to be mastered by the user to obtain the desired results. As a relevant example, air quality monitoring requires the integration of multiple data with different spatial and temporal granularities that are often distributed by more than one provider using not uniform formats and access methods. Besides traditional air pollution ground sensors observations, novel data sources have emerged. Among them, the Sentinel-5P mission of the European Copernicus Programme is one of the most recent Earth Observation platforms providing estimates of air pollutants with daily global coverage. These estimates are promising to foster air quality analysis and monitoring by complementing ground sensors observations. Therefore, the development of data handling and analysis strategies - allowing users for a smooth integration of satellite and ground sensor observations - is key to support future air quality studies. To that end, the present work investigates the use of the Open Data Cube as a single data endpoint to incorporate ground sensors and satellite observations into local air pollution analyses. A preliminary implementation is presented using the Lombardy region (Northern Italy) as a case study.

1. INTRODUCTION

Geospatial data has become a fundamental source of information for governments, companies and organizations. Its relevance mainly lies in the fact that it contributes to understanding better world phenomena and the consequences that humans activities have on Earth. A relevant example of the above, among many possible others, is the use of modern geospatial information for air quality monitoring (Faruque, 2019).

According to the European Environment Agency, every year more than 430,000 premature deaths European Union can be attributed to air pollution (EEA, 2020). The decline of air quality has pushed organizations such as the United Nations (UN), to envisage actions for a peaceful and prosperous development for future generations through the definitions of the Sustainable Development Goals (SDGs). Many SDGs stress the need to reduce the negative impact of air pollution (e.g. 3, 11 and 13 that addressed directly well-being, sustainable communities and life on earth) (United Nations, 2020). In this context, geospatial information support scientists to explore both effects and causes of air pollution as well as empower the identification of viable solutions. Significant public investments have pushed Geographic Information Systems Science (GIS-science) to retrieve key information to accomplish these SDGs. A significant example is the effort of national and international space agencies, including the National Aeronautics and Space Administration (NASA) and European Space Agency (ESA), in providing cutting-edge Earth Observation platform to support environmental monitoring, including air quality (Kansakar, Hossain, 2016).

Although geospatial data can nowadays provide scientists and policymakers with an outstanding amount and variety of information, there are still critical aspects that must be considered such as accuracy, spatial-temporal resolution, accessibility, and other factors that can affect the generation of impacting outcomes (Fowlie et al., 2019). Furthermore, a key concern is connected to the technical skills of end-users to proficiently take advantage of such information. This paper examines the example of air quality observations by outlining current usage drawback and identify Free and Open Source Software (FOSS) solutions to assist users in air quality observations handling tasks. Indeed, air quality observations are provided by many sources such as satellite ground sensors which require to be integrated for both global and local air quality analysis.

In this study, the Open Data Cube (ODC) data exploitation architecture (Killough, 2018) is applied to integrate Nitrogen Dioxide (NO_2) satellite estimates and ground sensors observations for the Lombardy region (Northern Italy; 23,844 km^2). The Lombardy region is one of the most densely populated areas in Europe and is heavily affected by air pollution. This area is a pollution hot-spot also due to its peculiar micro-climatic features, mainly characterized by wind channelling along the Po river valley, and frequent thermal inversions in mountain areas, which do not allow the correct dispersion of pollutants in the lower atmosphere (Diémoz et al., 2019). Air quality ground sensors in the region are managed by the Lombardy Region Environmental Protection Agency (ARPA Lombardia) while considered satellite estimates are the ones provided by the Sentinel-5P satellite of the European Copernicus Programme (de Vries et al., 2016).

The presented work exploits ODC to integrate into a single en-

*Corresponding author

dpoint both types of information thus leveraging their concurrent application into air quality monitoring, which is traditionally performed using ground sensor observations only, while exploitation of the modern satellite data assets has not been yet considered by the local authority.

The remainder of the paper includes an overview of the ODC software in Section 2. The description of data processing operations required by the ODC for the proposed data integration is reported in Section 3. The development of the proposed ODC application is included in Section 4. Finally, conclusions and future work are included in Section 5.

2. THE OPEN DATA CUBE

A relevant example of a FOSS solution for managing different data sources in a single tool is the ODC (Open Data Cube, 2021a). The ODC is available under Apache 2.0 license and it is tailored to access, manage, and analyse mainly satellite Earth Observation data. This software consists of a collection of Python tools acting as an intermediary layer between satellite data and the end-user. The main objective is to provide access to structurally complex files that alternatively would require high expertise from the user, meaning that a significant amount of time would be spent on data pre-processing instead of focusing on the data analysis. This software makes available to the user tools that can be used for data exploration, processing and analysis (see Figure 1). The main ODC tools and assets are summarized in Figure 1 and described in the following.

1. Command line tools: Available for developers to interact with the ODC to support the system architecture,
2. Open Data Cube Explorer: A web application that allows the user to interact with the satellite products available,
3. Open Data Cube Stats: Tool for statistical analysis of the products contained in the ODC,
4. Web User Interface: Interaction interface to communicate with the developed algorithms,
5. Jupyter Notebooks: Programming notebooks provided to actively interact with the algorithms,
6. Open Geospatial Consortium (OGC) Web Services: Connectors between non-OGC products and OGC products.

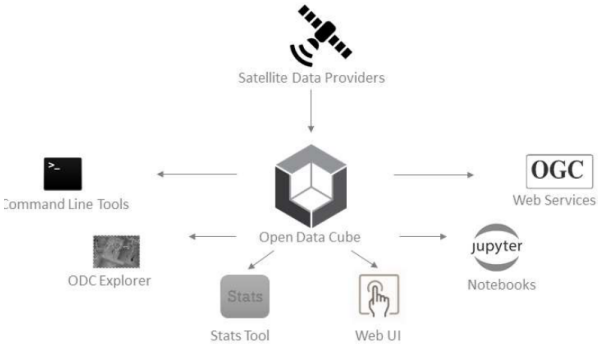

Figure 1. High level overview of the ODC tools (Open Data Cube, 2021a).

Existing implementations of the ODC are representative of the advantages that the ODC provides. The most mature deployments mentioned in the official ODC documentation (Open Data Cube, 2021b) are the Digital Earth Australia (https://www.ga.gov.au/dea), the Digital Earth Africa (https://www.digitalearthafrica.org), the Swiss Data Cube (SDC, https://www.swissdatacube.org, Figure 2) and Vietnam Open Data Cube (https://datacube.vn). It is worth mentioning that although these deployments are all based on the original tools designed for the ODC, each of them both focuses on different regions and consists of distinct approaches to the same set of tools, this means that the backbone of each deployment is the same, but the final product differs from one to another. Additional information on the features of the aforementioned ODC applications can be found on the official ODC website (Open Data Cube, 2021a).

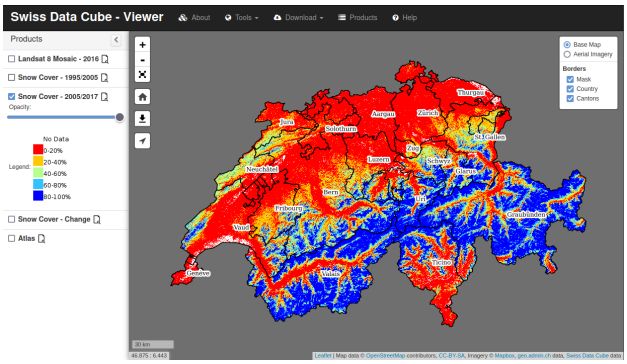

Figure 2. The SDC ODC Explorer (Open Data Cube, 2021a)

The current ODC implementation empowers end-users with tools for managing a number of well-established data products, including e.g Sentinel-2 and Landsat-8 satellite imagery, and derived products. Support to other data, including Sentinel-5P imagery and ground sensor air quality observations, is limited because built-in ingestion patterns for currently supported products cannot be directly reused. Basically, the additional development effort for integrating new data sources into the ODC consist of adapting raw data to formats (GeoTIFF or NetCDF) accessible by ODC core libraries (e.g. Python Rasterio, https://rasterio.readthedocs.io and XArray, http://xarray.pydata.org) and compiling metadata files for each of the products.

The ODC application proposed in this work aims at delivering data ingestion patterns that can be used to address specifically the Sentinel-5P air pollutants satellite estimates as well as ground sensors observations that are here intended as georeferenced time series in tabular format (e.g. CSV) which is a common exchange format for air quality ground sensor observations.

3. DATA DESCRIPTION

In this section, a description of both satellite and ground sensors data used for the proposed integration into the ODC is provided. The identified data pre-processing tasks are outlined accordingly.

3.1 Air pollution monitoring from space

Concerning air pollution, the most recent mission providing satellite estimates of tropospheric pollutants is the Sentinel-5P which carries on board the TROPOMI (TROPOspheric Monitoring Instrument), a multi-spectral imaging spectrometer, developed in a joint venture between the Netherlands Space Office and ESA (de Vries et al., 2016). Pollutants monitored by

the Sentinel-5P include Carbon Monoxide (CO), NO_2, Sulphur Dioxide (SO_2), Ozone (O_3), Methane (CH_4), Formaldehyde ($HCHO$), and aerosols. These measurements are performed globally at a spatial resolution of 5km x 3.5km and a daily temporal resolution. The data provided is regularly gridded and supplied in NetCDF format. This format allows the user to incorporate the estimates into the ODC without the need of converting the format. the Sentinel-5P data are open and can be downloaded through the Copernicus Open Access Hub (https://s5phub.copernicus.eu).

The Sentinel-5P Level-2 products - as downloaded from the data provider portal - requires to be resampled on a fixed grid in contrast to, for example, Sentinel-2 images which are already projected onto a UTM grid at Level-2. This prevents the direct ingestion of the product into the ODC by using the build-in functionalities of the current implementation.

3.2 Air pollution monitoring from the ground

The second data source used in this work is ground sensor air pollution measurements from ARPA Lombardia which provides continuous measurement of air quality used by the authority to compute indicators specified by current legislation (World Health Organization, 2006, European Commission, 2008).

The air pollution ground sensors network of the Lombardy region is composed by 85 fixed stations (see Figure 3) continuously monitoring concentrations of SO_2, NO_2, CO, Particulate Matter (PM_{10} and $PM_{2.5}$), O_3 and Benzene within the region. The observations are distributed through three channels. The first two channels are the ARPA Lombardia Web portal (www.arpalombardia.it) and the Lombardy region Open Data Portal (www.dati.lombardia.it), where data can be downloaded in CSV format. Lastly, data can be also accessed through a dedicated Application Programming Interface (API, dev.socrata.com/foundry/www.dati.lombardia.it/nicp-bhqi).
All the channels provide the measurements as georeferenced time-series of air pollutant concentrations with a temporal resolution of 1 hour.

Currently, the ODC is not meant to ingest geolocalized sensors observations in their native format (CSV). Therefore, ARPA data has to be transformed to make it compatible with the Python libraries that are used in the ODC and as mentioned before in Section 2. Sensors are scattered points while gridded data is requested by the ODC data ingestion procedure. To that end, the pre-processing for ground sensors observations developed in this work is designed to address this requirement and provide layered grids suitable for comparison and integration with the satellite products.

4. SATELLITE AND GROUND SENSOR DATA INTEGRATION

Air quality monitoring requires the use of a variety of information (e.g. satellite and ground observations) to analyse pollutants spatial-temporal patterns (Heal et al., 2012). Therefore, data integration is critical to air quality monitoring along with the availability of data access tooling for the end-users.

The exclusive use of ground sensors for air quality monitoring, source apportionment, and exposure modelling brings relevant limitations (Kansakar, Hossain, 2016, Fatimah, 2016). The main one is the limited spatial granularity and coverage of the measurements which are representative only of the neighbouring areas of the sensor network

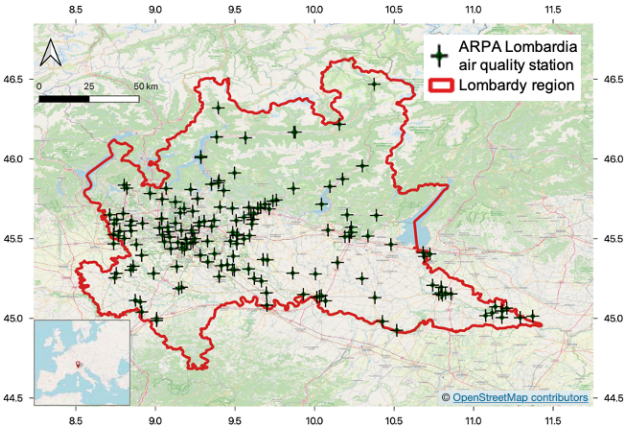

Figure 3. Locations of ARPA Lombardia Air Quality stations (Oxoli et al., 2020)

Satellite data provide instead global spatial coverage but - currently - has limited time granularity and requires to be complemented with ground sensors data because they are estimates derived from spectral analysis of imagery and not actual concentration measures (Fowlie et al., 2019). Therefore, air quality indicators cannot be directly computed from satellite observations by making their integration with ground observation a key issue to be addressed.

4.1 Data pre-processing

In the presented case study, satellite and ground sensors NO_2 data from January 1^{st} 2020 to April 14^{th} 2021 were considered. The pre-processing procedure that is specific for two data sources is detailed below.

4.1.1 Sentinel-5P Sentinel-5P data can be integrated into the ODC with a minimum pre-processing. The approach applied in this work was to transform the highest-level product available (Level-2) in the provider data portal to be subsequently refined to Level-3. Level-2 data consists of geolocated column measurements of the aforementioned pollutants and indication of the quality measured at a pixel level among other parameters. The conversion to Level-3 consists of several data transformations that optimize the datasets that will be contained in the ODC. The operation can be directly achieved by using the HARP tool provided by the ESA Atmospheric toolbox (https://atmospherictoolbox.org/harp) that implements user-ready libraries to manipulate satellite data specifically created for Sentinel products.

The first step is to clip the data on the desired area of interest (AOI) avoiding to load the complete tile (half orbit) of the satellite but a small portion of the acquired image. Additionally, a filter is applied using the Quality Assurance value (qa-value). In this study, the minimum quality value accepted for any pixel is 75% to ensure the removal of cloudy scenes, snow and ice-cover and problematic retrievals that may be present in the images. Finally, the bands are extracted. This is a procedure that reduces the occupied memory space by more than 95% in the in case of the tropospheric NO_2 product for the Lombardy region (i.e. 400 MB for Level-2 full NetCDF and 25 MB for the Level-3 processed NetCDF).

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVI-4-W2-2021-31-2021 | © Author(s) 2021. CC BY 4.0 License.

4.1.2 Air quality ground sensors Ground sensor air quality measurements provided by the ARPA Lombardia network were pre-processed following the steps described in the following. The pre-processing for these observations is more intensive than the Sentinel-5P due to their native format which is not currently supported by ODC build-in data ingestion functionalities, as explained in Section 3.2.

1. Data procurement: ARPA Lombardia Air quality ground sensors data are provided in tabular format and were downloaded through the API using Python. Data from years previous to the current one (2021) are not served by the API and had to be downloaded in CSV format manually. The Pandas Python library (https://pandas.pydata.org) was used to merge the whole files into a single table.

2. Data gridding: ODC ingestion requires data to be on a regular spatial grid. This is imposed by the formats that the ODC system uses (i.e. NetCDF or GeoTIFF). A grid with a resolution of 0.01 degrees was selected according to the minimum distance between all of the sensors. Once the grid was created, sensor observations were assigned to each pixel centre using the Nearest Neighbours interpolation, appointing Null values to those pixels that do not overlap with any sensor location.

3. Data formatting: Once the data is re-gridded it was exported to NetCDF. Despite being required by the ODC, this format allows multi-layering of the information that was exploited to stack multitemporal observations from each sensor which constitutes the data cube. The peculiarity of this step is that each day is stored in an independent file, replicating the structure of satellite data. This implied extracting sensor observations at the passage time of the satellite (around 12:00 UTC for the Sentinel-5P) to obtain a grid for each day of the time period to be compared with the one provided by the satellite.

4.2 Integration into the Open Data Cube

Once the Sentinel-5P and ARPA Lombardia files have been pre-processed they were input to the ODC system.

The ODC works by employing processing and visualization tools that help the user access and manipulate geospatial data. It is important to emphasize that the ODC is not a storage system, instead, it is a platform that manages information contained in a file system or database, acting as an intermediate layer between the user and the files. For this reason, the ODC data integration consists of the indexing of NetCDF or GeoTIFF files in the database by means of the metadata generation for each of the target data file.

Data indexing into the ODC database consisted of two steps, explained below, both depending on metadata files in YAML format. These files are structured in EO3 format and help the software understand how to manage and group each of the files into a single-layered product that contains information on the available data measurements, platform sensor and names, geospatial extent and projection, acquisition time and other metadata specified in the ODC documentation (Open Data Cube, 2021b).

1. Product definition: The first step for the ODC indexing consisted in creating a group in the database called *product*. This group helps the system to identify files belonging to a specific data source (e.g. satellite or ground sensor) and a specific pollutant. Therefore two products were created which identify both the Sentinel-5P and the ARPA Lombardia NO_2 data. These products were further referenced in the dataset indexing and in the measurements extraction.

2. Dataset indexing: The final step for the data integration in the ODC was to register each of the files in the indexed database. In this part of the procedure, each NetCDF files gets ingested into the ODC system to be extracted in future steps. This consists of registering into the system a YAML file that contains both measurements and product information and it is used to group files into a single product.

The generation of the YAML files for each NetCDF is a manual task that is prone to mistakes and can consume a large amount of time. For this reason, automation is a fundamental step for data indexing. The creation of YAML files and their association to each NetCDF was implemented by means of a Python script that read each of the available NetCDF files to extract its properties and transcribe them into a template YAML file and finally runs the ODC command to ingest the NetCDF and the associated metadata.

4.3 Data integration results and application

The result of data pre-processing and integration can be accessed through any exploratory data analysis Python tool that can be run programmatically using live scripting e.g. in a Jupyter Notebook.

A first approach to the data exploration concept is depicted in Figure 4, where the Juypter Notebook is used to explore the ground sensors product for NO_2 (called here *ARPA_NO2*). Figure 4 shows how the XArray Python library can be used for extracting data according e.g. to their acquisition time and coordinates that are used as indexes for accessing the ODC products content.

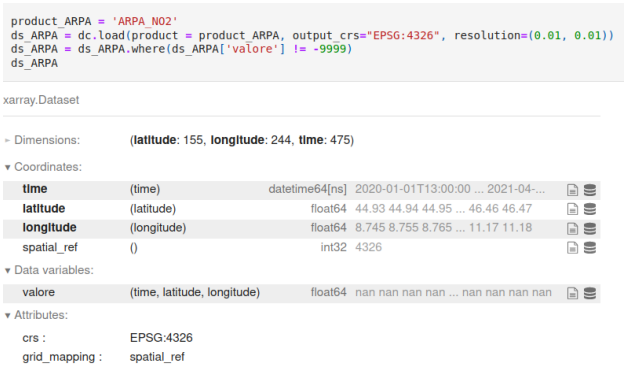

Figure 4. Example of the extraction of ARPA Lombardia NO_2 ODC product using the XArray library in a Jupyter Notebook

Further than data exploration, an example of correlation analysis between the ARPA Lombardia and the Sentinel-5P data was performed by leveraging exclusive the generated ODC products as follows. Both datasets were extracted using XArray and overlaid. Data for the whole analysis period (January 1^{st}, 2020 to April 14^{th}, 2021) were considered in this test. The Pearson's correlation coefficients between NO_2 satellite estimates and ground sensors observations at each sensor's location

were computed using Pandas. Results are reported in Figure 5 and provide insight on the quantitative comparability of the two air quality observations (Oxoli et al., 2020) by suggesting the ODC as a valuable tool to perform such investigations.

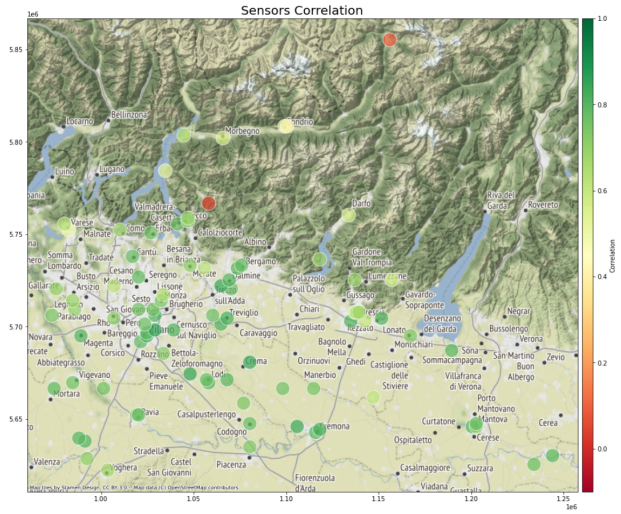

Figure 5. Map of Pearson's correlation coefficients computed between the time-series of NO_2 observations from the sensors and the Sentinel-5P at each ARPA Lombardia sensor location in the Lombardy region.

The outcome of the presented work demonstrates that integration of new data sources into the ODC is possible, although a pre-processing procedure is needed to comply with the system requirements. The designed data ingestion procedure is promising to extend the capabilities of the ODC to data different from satellite Earth Observation.

The proposed data integration of Sentinel-5P and ARPA Lombardia ground sensor data is critical for the considered case study on air quality monitoring. The enabled possibility of concurrent access and analysis of the two types of observations through a unified data endpoint opens newsworthy possibilities to scientist and local authorities to unpin the capabilities of cutting-edge informational assets - such as the one provided by the modern EO platforms - alongside their operational routines based on traditional ground sensors data.

5. CONCLUSION AND FUTURE WORK

This paper presented a procedure to integrate heterogeneous air quality observations, such as satellite and ground sensors data, into the ODC system. The integration was tested in support of the local air quality monitoring in the Lombardy region, with the aim of demonstrating the advantages of the ODC approach for enabling ground and satellite-based information being synergically employed to foster air quality monitoring, studies and policy-making by demanding to end-users less specific skills for data access and manipulation.

Data integration is one of the current challenges when dealing with geospatial data, especially because data heterogeneity (i.e. coming from different sources in different formats) require the users to understand the delivered data by the different providers increasing the analysis complexity. To that end, the ODC is identified as a state-of-art support system to deliver analysis-ready data. The ODC system has currently been adopted operationally by few organizations, and it still doesn't incorporate most of the satellite data and none of the non-satellite data sources. However, the system can be adapted to these data through automatic procedure tailored to different data integration needs.

Future work will focus on ingesting into the ODC other air quality measurements that have not been considered in this work. In parallel, the operational application of the proposed tools and data will be investigated through the interaction with local stakeholders, including ARPA Lombardia. Questions regarding the computing infrastructure to support both the development and publication of the ODC instance will be also addressed to guarantee the end-users access to a large amount of information and analysis tools remotely.

REFERENCES

de Vries, J., Voors, R., Ording, B., Dingjan, J., Veefkind, P., Ludewig, A., Kleipool, Q., Hoogeveen, R., Aben, I., 2016. Tropomi on esa's sentinel 5p ready for launch and use. *Fourth International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2016)*, 9688, International Society for Optics and Photonics, 96880B.

Diémoz, H., Barnaba, F., Magri, T., Pession, G., Dionisi, D., Pittavino, S., Tombolato, I. K., Campanelli, M., Ceca, L. S. D., Hervo, M. et al., 2019. Transport of Po Valley aerosol pollution to the northwestern Alps–Part 1: Phenomenology. *Atmospheric Chemistry and Physics*, 19(5), 3065–3095.

EEA, 2020. Premature deaths attributable to air pollution. *European Environmental Agency*. https://www.eea.europa.eu/media/newsreleases/many-europeans-still-exposed-to-air-pollution-2015/premature-deaths-attributable-to-air-pollution.

European Commission, 2008. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. *Official Journal of the European Union*.

Faruque, F. S., 2019. Geospatial technology in environmental health applications. *Environmental monitoring and assessment*, 191(2), 1–6.

Fatimah, D. Q., 2016. Variation in global chemical composition of PM2. 5: emerging results from SPARTAN. *Atmospheric Chemistry and Physics (ACP)*, 16(15).

Fowlie, M., Rubin, E., Walker, R., 2019. Bringing satellite-based air quality estimates down to earth. *AEA Papers and Proceedings*, 109, 283–88.

Heal, M. R., Kumar, P., Harrison, R. M., 2012. Particles, air quality, policy and health. *Chemical Society Reviews*, 41(19), 6606–6630.

Kansakar, P., Hossain, F., 2016. A review of applications of satellite earth observation data for global societal benefit and stewardship of planet earth. *Space Policy*, 36, 46–54.

Killough, B., 2018. Overview of the open data cube initiative. *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 8629–8632.

Open Data Cube, 2021a. The open data cube. https://www.opendatacube.org. Accessed 29.05.2021.

Open Data Cube, 2021b. Open data cube manual. https://datacube-core.readthedocs.io. Accessed 29.05.2021.

Oxoli, D., Jimenez, J. C., Brovelli, M., 2020. ASSESSMENT OF SENTINEL-5P PERFORMANCE FOR GROUND-LEVEL AIR QUALITY MONITORING: PREPARATORY EXPERIMENTS OVER THE COVID-19 LOCKDOWN PERIOD. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 44(3/W1).

United Nations, 2020. The sustainable development goals report 2020. https://sdgs.un.org/goals. Accessed 29.05.2021.

World Health Organization, 2006. Air quality guidelines: global update 2005: particulate matter, ozone, nitrogen dioxide, and sulfur dioxide.