Je-LKS
Journal of e-Learning and Knowledge Society
The Italian e-Learning Association Journal
PEER REVIEWED PAPERS
STEPS TOWARD THE DIGITAL AGENDA: OPEN DATA TO OPEN KNOWLEDGE
Vol. 10, n.2, 2014
ISSN: 1826-6223 | eISSN: 1971-8829

LOMBARDIA EPA OBTORTO COLLO DATA AND ANTI-POLLUTION POLICIES FALLACIES

Andrea Trentini
Dipartimento di Informatica, Università degli Studi di Milano
Milano (Italy)
andrea.trentini@unimi.it

Keywords: open data, public accountancy, particulate matter, anti-pollution policies, web scraping

ARPA Lombardia EPA[1] maintains pollution monitoring stations since 1999. Measured data is "somewhat" available through ARPA's website. "Somewhat" because a CAPTCHA protected download request form must be filled up for every combination of (station, pollutant, timeframe $<$ 1yr). I.e., citizens wanting the whole dataset should complete thousands requests by hand. In 2003 Lombardia government decided to ban classes of vehicles due to a supposed air pollution increase. The author searched for data to counteract that decision - air pollution is in fact almost constantly decreasing since the seventies - found about ARPA's overcomplicated download procedure and downloaded some of the data, just the amount needed to argument against the ban. He also decided to set up an automated system to overcome that difficulty for future purposes. The system is working since 2005. The whole set of collected data has been published (*freed*) on the

[1] Environment Protection Agency

for citations:
Trentini A. (2014), *Lombardia EPA obtorto collo Data and anti-pollution policies fallacies,* Journal of e-Learning and Knowledge Society, v.10, n.2, 101-111. ISSN: 1826-6223, e-ISSN:1971-8829

author's website, thus implementing the title's *obtorto collo* (from latin, it means "reluctantly", "unwillingly") data from ARPA. This paper describes ARPA barriers (1), the automated collecting system (3.1), data elaboration (3.2) and some of the findings based on that data (3.3).

1 Introduction

Air pollution started to become a problem with the industrial revolution (Seinfeld & Pandis, 2012; Lave & Seskin, 2013; Steinle *et al.*, 2013). During the twentieth century pollution skyrocketed to the extension that some notable high *peaks* were even given a name such as the "Great Smog of '52" (Davis, 2002). In the seventies governments started to legislate (US-EPA, 2013) to try reducing emissions. From then on, air pollution slowly began to decrease as new generations of technologies replaced older ones (see Figure 1 and similar graphs[2] available on ARPA).

A typical example context is the set of land transportation technologies (i.e., vehicles) we use every day to commute, travel, have fun and so on. Since the original "Clean Air Act" (US-EPA, 2013), cars, motorbikes, buses, etc. makers have been compelled to fulfill ever updated emission requirements. Stricter rules substitute older ones as technologies progress. Europe has followed this trend with the so-called EuroX legislation (European Commission, 2014) to impose maximum emission limits for every type of car+engine (gasoline, diesel, 4-stroke, 2-stroke, hybrid-electric, etc.) produced and sold. E.g., for passenger vehicles (cars), EuroX rules define the following pollutants that should be regulated: *CO* (Carbon Monoxide), *THC* (Hydrocarbon), *NMHC* (Non-methane hydrocarbons), NO_x (Nitrogen oxides), $HC+NO_x$, *PM* (Particulate Matter), *Px* (Particle number, this last one is still in the process of being detailed, it is not part of the rules). It is also important to note (it will be useful later) that EuroX rules do not impose limits to *PM* for gasoline vehicles, since gasoline (4-stroke) engines do not produce any *PM* (Schauer *et al.*, 1996) for practical purposes. Moreover, many countries declared laws - e.g. Europe (European Commission, 2008) - to limit pollution concentration in the air. Europe legislation was adopted in Italy with the following upper bounds, here associated to the measured (see tables in section 3) average in 2013: SO_2<125μg/m^3 → **<5**[3] (very low); *PM10*<50μg/m^3 → **38** (almost low); *PM2.5* is not yet specified but already monitored since EU commission is still debating about it, the proposed limit is 20 → **29** (high); NO_2<200μg/m^3 → **86** (low); *CO*<10μg/m^3 → **1.5** (very low); O_3<180-240μg/m^3 → **30** (very low); Benzene (no bounds), → **1.4**. *PM10* is also monitored in terms of "number of limit excesses" during the year, and the number of excesses should remain under a fixed number (usually 35/yr). This

[2] http://ita.arpalombardia.it/ITA/qaria/img/qaria/graficiInqNew/<prov>_<pollut>.png such as http://ita.arpalombardia.it/ITA/qaria/img/qaria/graficiInqNew/MI_PM10.png

[3] Monitoring stations print this value under low concentration conditions

is because *PM10* is a very cyclical pollutant: increasing in winter and decreasing in summer. Currently, here, the *PM10* average is below the limit, but the number of excesses/yr exceeds the EU prescription.

Summing up: almost every monitored pollutant is below the upper limit and the only one that should be taken into account is Particulate Matter (*PM10* and *PM2.5*).

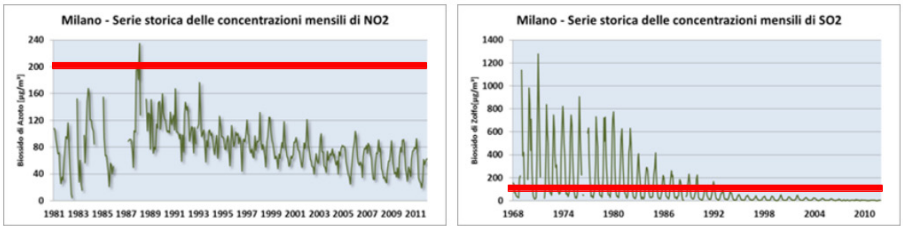

Fig. 1 - NO_2 (since 1990), SO2 (since 1970) with upper bounds (red line), source: ARPA

Around the year 2000, despite the downward trends in air pollutants, some Italian local governments started introducing legislation restricting the use of private vehicles due to a supposed "upward trend" in pollutants. E.g., in 2003 and 2006, the Lombardia government decided to ban some classes of old vehicles (regardless of engine type) due to a supposed air pollution increase (!). Many vehicle owners could no longer drive their cars, motorbikes, etc. while still paying ownership taxes and mandatory insurance. Even in case of 4-stroke gasoline vehicles which do not produce *PM10* and *PM2.5*. While public transport buses with very old (average public vehicles age is about 20 yrs in Italy) diesel engines could pollute without any limitation. The difficulty to understand (engine, pollution, etc.) technology (Schwarz & Thompson, 1990) is probably the source of this "unreasonable" ban. In Italy, the "principle of law reasonability and proportionality" (Enciclopedia Treccani, 2014; Cerri, 1994) is one of the most important principles used to (in)validate laws, it states that: "*The reasonability principle (RP) is a corollary of the principle of equality, drafted by the Constitutional Court, inspired by a similar principle identified by the Anglo-Saxon jurisprudence. The RP requires that the provisions contained in acts having the force of law must be appropriate or congruent to the objective pursued. It is therefore a breach of RP when there is a major contradiction within a law, or between it and the public interest pursued. The RP is therefore a limit to the discretion of the legislature, which prevents arbitrary exercise. The verification of RP of a law involves t****he investigation of its assumptions of fact****, the assessment of the* ***congruence between means and goals*** *... In case*

it is established the irrationality of the law, it will be affected by the vice of ***excessive legislative power****, and, as such, can be held to be unconstitutional by the Constitutional Court ...*".

Of course the ban proposal worried many citizens to no small end. In fact, the author of this paper, when reading the proposed banning rules and, above all, the motivations (the aforementioned supposed air pollution increase), began searching for information about air pollution and technical data to write an "open letter to the administration" that circulated on media and among citizens. He found out about EuroX legislation, about what other countries did (almost never permanent bans, and banned vehicles could always be upgraded to cleaner ones with aftermarket components, something not allowed by present Italian legislation) and about ARPA air monitoring. The letter disclosed the many incongruosnesses in the proposed ban and brought some technical awareness in politicians and citizens: eventually the Lombardia ban was softened, i.e., applied to very old vehicles and to 2-stroke motorbikes only.

ARPA Lombardia EPA (Environment Protection Agency) maintains a network of pollution monitoring stations from downtown Milano to remote places near the mountains since 1999. When the author found the ARPA website[4] he searched for a way to download historical data, to show that the "supposed" increase in air pollution advocated by the local administration was in fact a decrease. In fact the data were there, available for download and in proper format: CSV (Comma Separated Values) files, i.e., 3 stars (Berners-Lee, 2009a) OpenData graded. But with some "webstacle" (web+obstacle crasis) to impede full data exploitation. If a citizen wanted to collect one set of data he/she should fill and submit a web form, then data are sent by email (i.e., the procedure is not completely anonymous). Moreover, the citizen can request only a subset of data for each form submission, i.e.: one single station, one single pollutant, one single timeframe smaller than one year. Summing up, if a citizen needs the whole dataset he/she should prepare him/herself to complete about 80 (stations) x 7 (pollutants per station) x 15 (years of monitoring) requests, by hand since the download request web form is CAPTCHA[5] protected.

This is the motivation for the author's decision to create an automatic web grabbing system. The system was then developed and it is working almost continuosly since 2005. It is described in 3.1 and 3.2.

In recent years (starting around 2008) the Milan City Council started using very similar methods as those used by the Lombardia local government in 2003 to justify introducing a congestion charge, first called "Ecopass" and then "Area C", in an 8 km^2 area roughly corresponding to that inside the XVI century Spanish Walls. The name change is very telling, as it was introduced

[4] http://arpalombardia.it

[5] Completely Automated Public Turing test to tell Computers and Humans Apart

after its air-pollution reducing effectiveness was debated by media and citizens. Still, the legal (and political) justification of "Area C" is heavily based to this day on the idea it's used to fight pollution (http://www.areac.it). This is due to the fact that Italian laws only allow traffic restrictions of this kind in case of a well documented and serious threat to the public health, see (Ministero delle Infrastrutture e dei Trasporti, 2010).

2 *Obtorto collo* data

Open Data has become a worldwide movement involving governmental and non-governmental actors. The Open Knowledge Foundation (OKF) was one of the first organizations to define "openness" in this context and it has recently given birth to (Open Knowledge Foundation, 2014) to formalise meta-knowledge about open knowledge. The OKF definition of "openness" can be quoted as: "A piece of data or content is open if anyone is free to use, reuse, and redistribute it - subject only, at most, to the requirement to attribute and/or share-alike". Moreover, Tim Berners-Lee (Berners-Lee, 2009a) defined a five star rating for Open Data to highlight the importance of not just legal but also technical aspects of openness, for example through the use of open standards and non-proprietary file formats for Open Data publishing. More broadly, Berners-Lee and others (Berners-Lee, 2009b; Bizer *et al.*, 2009) promoted the concept of Linked Open Data to transform "data on the web" into "the web of data" by encouraging the linking of one's own data with other datasets. HM Government's Open Data White Paper (HM Government, 2012) states that Open Government Data is "Public Sector Information that has been made available to the public as Open Data" and defines Public Sector Information (PSI) as "data and information produced, collected or held by public authorities, as part of their public task", data that should be accessible (ideally via the internet) at marginal cost and without discrimination, available in digital and machine-readable format, and provided free of restrictions on use or redistribution.

The author believes that since ARPA Lombardia is government, it is in fact a public agency owned by Regione Lombardia, it should publish all the data without any webstacle so that citizens, ONGs, media, etc. could easily get them, analyze them and publish their studies to compare with governmental claims.

Obtorto collo comes from latin, it means "reluctantly", "unwillingly". ARPA decided not to let easy access to their data but someone, the author, overcame their barriers and "freed/opened" scraped data on his personal website (http://arcipelagoareac.it) under a Creative Commons license. That is why the author proposes a new category name for data besides "free/open" and "proprietary/closed": ***obtorto collo* data**, which means "intended to be closed but opened by a third party" (see also section 4 for an interesting anecdote).

3 Methodology

This section summarizes (due to space constraints) the "scrape" (collect from web), store and analyse ARPA data process. *Webscraping* is a common technique to extract (potentially massive) data from websites in a programmative manner (instead of "by hand"). The author developed his system as a set of BASH scripts running under Debian GNU/Linux with the following modules: ARPA page gathering (download and store); data extraction (from stored pages), based on Unix filters to generate parsable data (i.e., CSV files); data analysis. The GPL licensed scripts (and data, and graphs) are downloadable from author's site http://arcipelagoareac.it.

3.1 Data scraping

Inspiration came from archive.org WayBackMachine, a freely available service saving web pages snapshots for "trusted citation in the future". The author wrote a simple script to get periodic snapshots of a subset of ARPA website. Since ARPA dataset download form cannot be submitted automatically, the author used a page that displays current data day by day, for any given Lombardia subregion. It was easy to reverse engineer the URL parametrization, and to update it when ARPA changed website structure. The script is based on wget and it ran every night for about ten years. The script saves every subregion page http://ita.arpalombardia.it/ITA/qaria/listaXX.asp, (the numbers *XX* correspond to areas in our region, e.g., 01 for Milan, etc.) as a compressed html.gz file with a full date filename.

3.2 Data cleaning and verification

Author's goal was to create CSV files, aggregated by monitoring station, from the set of html day-by-day saved pages. Data extraction and cleaning is a bit complex, ARPA web pages are not W3C[6] valid, a page excerpt:

```
...class="linktesto"><B>Cormano</B><br><font color="#666666"></font></a></td><td width="1" bgcolor="#003366"><img border="0" src="img/qaria/pxb.gif" width="1" height="28">...</td><td width="54" align="right" class="testobox">86</td>...
```

Relevant info (i.e., data to be extracted) is in the first ("Cormano") and in the last ("86") lines. In this case tools like `tidy` (a validator/indenter/cleaner), `html2` (a tool from the `xml2` package) or other well-formed html expecting tools fail or behave erratically. A trial and error combination of Unix filters

[6] http://validator.w3.org

achieved an acceptable data extraction, here are some script excerpts[7]:

```
NOME=$(grep -v html $FILE | sed 's/<br>/ /g' |\
  vilistextum -a - - | grep 'Dati' | tr -s " " |\
  cut -f4- -d" " | awk '{print $3"-"$2"-"$1}') ...
echo $linea | egrep -qi "(legenda|stazione)" ...
echo $linea | vilistextum -a - - | tr "\n" " " ...
done | sed -f sostituzioni.sed |\
  sort -n -t "/" --key=1 --key=2 -key=3 |\
  tr -d " " >> $OUT
```

These Unix filters are common tools: `(e)grep`, to select lines in a file based on pattern; sed, to substitute strings; `awk` and `cut`, to extract fields; tr, to substitute chars and `vilistextum`[8] "... a html to text/ascii converter specifically programmed to get the best out of incorrect html". The procedure generates a set of CSV files with chronological data, e.g.:

```
Date,       SO2,PM10, PM2.5,  NO2,   CO,    O3,    Benzene
2005/12/5,   ,   ,     143,   3.2,    ,     6.8,
2005/12/6,   ,   ,     196,   4.6,    ,     9.0,     ...
2014/04/12,  ,   27,    18,    74,   1.4,    ,       <3
2014/04/13,  ,   28,    19,    71,   1.4,    ,       <3
2014/04/14,  ,   41,    32,   103,   1.5,    ,       <3
```

3.3 Analysis

The *Comune di Milano*, through its local EPA agency AMAT[9], claims (Bedogni, 2013b; 2013a) the following "successes": exhaust *PM10* = -58% (compared to 2010); total *PM10* = -40% (idem); Elemental Carbon = -61% (idem); Organic Carbon = -33% (idem); Ammonia = -48% (idem); NO_x, volatile organic matter, benzopirene = unquantified decrease (idem); CO_2 = -29% (idem); Methane = -19% (idem); NO_2 = -24% (idem) and an unspecified decrease of air pollutants **inside** AreaC **compared to** the area **outside**. Please bear in mind that these claims **are not based on air pollution measurement** but on computed figures, in fact AMAT declares[10] to use the COPERT (Katsis *et al.*, 2012) method to associate an "emissive weight" to every vehicle, then it "multiplies" that weight by the number of vehicles entering AreaC (!). Neither AMAT nor the Comune di Milano own/maintain permanent and EU certified air pollution measuring stations. While ARPA, of course, does.

[7] Just to give an idea of the script "complexity", please download the complete script archive for more information

[8] http://bhaak.dyndns.org/vilistextum

[9] Agenzia Mobilità Ambiente Territorio

[10] http://areac.amat-mi.it/it/areac/emissioni-da-traffico (in italian)

Table 1
MILANO SENATO YEARLY AVERAGES

Year.	PM10 (< 50)	PM2.5 (proposed < 20)	NO2 (< 200)	CO (< 10)	Benzene (unbounded)
2008	39.39	NA	93.50	0.86	1.43
2009	44.39	NA	113.66	1.40	2.10
2010	40.10	NA	100.71	1.51	1.20
2011	48.92	NA	100.01	1.61	1.44
2012	42.25	36.56	82.96	1.32	0.53
2013	38.03	29.28	86.27	1.49	1.40
2014	44.96	37.34	91.39	1.70	1.53

ARPA stations can generally monitor the following pollutants: SO_2, *PM10, PM2.5,* NO_2, *CO,* O_3, Benzene, but not every station can sense all the pollutants, e.g. station nr. 548 (see below) can measure just *PM10, PM2.5,* NO_2, *CO* and Benzene. Since the author got data from the ARPA network he willl argue only on the pollutants for which he has data, i.e., Elemental Carbon, Organic Carbon, Ammonia, CO_2 [not a pollutant but a "greenhouse gas" and it's not monitored neither limited by EU legislation (European Commission, 2008)] and Methane will be dropped from the AMAT list.

For the remaining pollutants we will mostly concentrate on the ones still exceeding the limits (see section 1), i.e., *PM10, PM2.5*. We will use data from one station inside AreaC to debate **trend claims** and we will compare data from two stations, one inside and one outside, to debate the **decrease claim** inside AreaC: **nr.548**, Milano Senato, inside AreaC, 1 km from the city centre, halfway to AreaC boundary, *PM10, PM2.5,* NO_2, *CO* and Benzene; **nr.531**, Limito Pioltello, outside AreaC (no ban on any vehicle), 12 km from the city centre, SO_2, *PM10,* NO_2, *CO,* O_3.

Data (http://arcipelagoareac.it/CSV [*]) and every generated graph (http://arcipelagoareac.it/Grafici) is available on the author's website.

[Trend claims] can be challenged looking at Table 1, generated from scraped ARPA data ([*] MilanoSenato.csv). PM10 in 2010 was almost the same as in 2013, 40.10 vs. 38.03, a 5%[11] decrease: definitely not a 58% ("exhaust") neither a 40% ("total") decrease. *PM2.5* measurement began in 2012 (and remained stable since then) so there is no reference for 2010. NO_2 was 100.71 in 2010 and 86.27 in 2013, a 15% decrease. CO was 1.51 in 2010 and 1.49 in 2013, a 2% decrease. Benzene was 1.20 in 2010 and 1.40 in 2013.

[11] Percentages are computed raw against the average, without σ normalization, to compare to percentage-measured claims.

Table 2[12]
PM10 YEARLY AVERAGES, IN/OUT

Year.	average PM10 in	average PM10 out	average of diffs [in-out], also in σ
2008	40.23	44.20	-4,94 (-0.18σ)
2009	44.33	48.99	-5.45 (-0.19σ)
2010	40.96	38.37	+3.38 (0.12σ)
2011	49.15	44.68	+4.55 (0.16σ)
2012	42.31	37.05	+1.38 (0.05σ)
2013	38.56	39.98	+2.15 (0.07σ)
2014	42.85	44.09	-1.56 (-0.06σ)

[In-area reduction claims] can be challenged looking at the average differences (*inside−outside*) between Milano Senato station (inside, with ban) and Limito Pioltello station (far outside, no ban), they do not in fact show evident signs of effect. Yearly differences (Table 2, generated using [*] MilanoSenato.csv and LimitodiPioltello.csv data files) and monthly ones (Figure 2) show that the "effect" of banning vehicles in Milano city centre may weigh between about +0.2σ (in 2011) and about -0.2σ (2009) - data σ is ~28 - and a question also arises: from 2008 to 2013 forms of vehicle banning were always in place, so why the difference is so low and rippling?

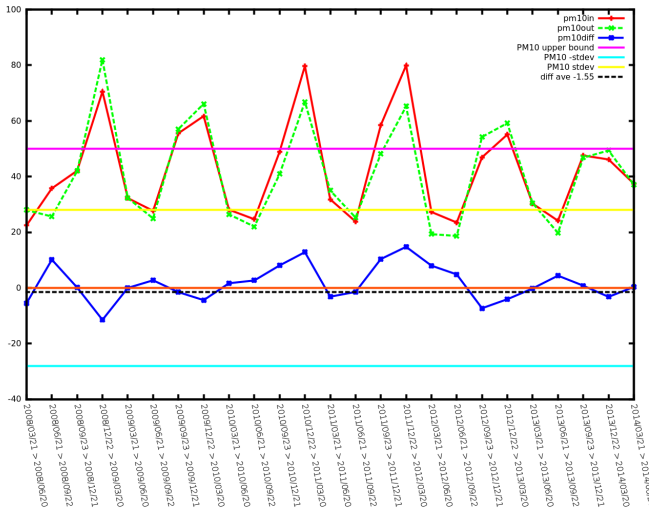

Fig. 2 - Values and difference seasonal averages

[12] Column 4 of the table - Average of diffs: Not the difference of other columns but the real average of the whole (in-out) set

Conclusion

This article describes author's work to extract ARPA air pollution monitoring data. Data gathered from 2005 through 2014 were useful to argument against misconceptions-based vehicle banning laws. Collected data have been made freely available as CSV files on the author's website. Data show that the two main Milano (Italy) bans, "Ecopass" (2008) and "AreaC" (2012), had undetectable effects on air pollution. The conclusion is consistent with the London Transport technical report (Transport for London, 2014) about the London congestion charge, similar to AreaC: "*Even so, trends in actual measured air quality continued to primarily reflect the diversity and dominance of external factors in determining pollution concentrations and, as such, did not allow the identification of a clear 'congestion charging effect'. ... Despite substantial reductions to road traffic emissions in London, trends in measured air pollution remain broadly static.*". The author has recently presented his work at the International Open Data Day (Italy) where a representative of Lombardia Informatica (LI, the ITC division of Lombardia government) was also invited to speak about Lombardia Open Data portal[13]. The event was also tweeted to the extent that a worried LI representative supervisor called him after the event to ask about the author's work (the representative then called the author as a kind FYI) and to say that they are in the process of changing ARPA data access policies. The author believes that this event alone was worth the effort, a political success.

REFERENCES

Bedogni, M. (2013a), *Emissioni atmosferiche da traffico stradale a milano - periodo gennaio - aprile 2013* (in italian). Technical report, AMAT-MI.

Bedogni, M. (2013b), *Emissioni atmosferiche nella città di milano - periodo gennaio - giugno 2013* (in italian). Technical report, AMAT-MI.

Berners-Lee, T. (2009a), *Linked data*. http://www.w3.org/DesignIssues/LinkedData.html.

Berners-Lee, T. (2009b), *Linked data*. http://w3.org/DesignIssues/GovData.html.

Bizer, C., Heath, T., and Berners-Lee, T. (2009), *Linked data-the story so far*. International journal on semantic web and information systems, 5(3):1–22.

Cerri, A. (1994), *Ragionevolezza delle leggi*. Enciclopedia giuridica, 25.

Davis, D. L. (2002), *A look back at the london smog of 1952 and the half century since. Environmental health perspectives*, 110(12):A734.

Enciclopedia Treccani (2014), *Ragionevolezza delle leggi* (in italian).

[13] http://dati.lombardia.it, it does not include ARPA data.

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.

European Commission (2014). Transport and environment - road vehicles - legislation.

HM Government (2012), *Open data white paper, unleashing the potential*. Technical report.

Katsis, P., Ntziachristos, L., and Mellios, G. (2012), *Description of new elements in COPERT 4 v10.0*. http://www.emisia.com/files/COPERT4_v10_0.pdf.

Lave, L. B. and Seskin, E. P. (2013), *Air pollution and human health*, volume 6. Routledge.

Ministero delle Infrastrutture e dei Trasporti (2010), *Codice della strada*, articolo 7: Regolamentazione della circolazione nei centri abitati (in italian).

Open Knowledge Foundation (2014), *History of the Open Definition*.

Schauer, J. J., Rogge, W. F., Hildemann, L. M., Mazurek, M. A., Cass, G. R., and Simoneit, B. R. (1996), *Source apportionment of airborne particulate matter using organic compounds as tracers*. Atmospheric Environment, 30(22):3837–3855.

Schwarz, M. and Thompson, M. (1990), *Divided we stand: Redefining politics, technology and social choice*. University of Pennsylvania Press.

Seinfeld, J. H. and Pandis, S. N. (2012), *Atmospheric chemistry and physics: from air pollution to climate change*. John Wiley & Sons.

Steinle, S., Reis, S., and Sabel, C. E. (2013), *Quantifying human exposure to air pollution—moving from static monitoring to spatio-temporally resolved personal exposure assessment*. Science of the Total Environment, 443:184–193.

Transport for London (2014), *Congestion charging publications*. http://tfl.gov.uk/roadusers/congestioncharging/6722.aspx.

US-EPA (2013), *History of the clean air act*. http://www.epa.gov/air/caa/amendments.html.

Wikipedia (2014), Web scraping - wikipedia, the free encyclopedia. [accessed 2014].