



Filtering data: Exploring the sociomaterial production of air

Vanessa Weber

abstract

Inspired by contemporary visions of the omniscience of data, this note questions the idea that data are raw, untreated, numeric replications of the world, available for immediate use at any particular moment in time. Instead, I argue that data are produced by different sociomaterial practices and therefore are neither raw replications of the world nor available instantaneously. Based on ethnographic research on environmental monitoring, this note explores the practices of filtering massive amounts of environmental sensor data produced in European cities in recent years. First, it reveals how the ongoing efforts to produce environmental data on air quality are enabled through increasing implementation of sensor technologies in the urban realm, whereby the assemblages of human and nonhuman actors and their interplay are unfolded. Second, the examination of the sociotechnical ‘sensing’ of air quality within different spatiotemporal techniques of filtering inevitably casts doubt on the distinction between the terms ‘data’ and ‘information’.

Introduction

At first we feel nothing, we are insensitive, we are naturalized. And then suddenly we feel not something, but the absence of something we did not know before could possibly be lacking. (Latour, 2006: 105)

Although air is a necessary and substantive part of life – it surrounds us; it flows into us and out of us – most of the time we are unaware of its existence.

Even though we generally know air's chemical composition and some of us measure its humidity and temperature in our living rooms, we almost never perceive it. Thus, the possibility of actually perceiving air may be understood as inherently intertwined with the technical infrastructures that produce our relation to it. However, air still remains abstract to us, something that cannot be grasped, something universal and yet not present, as Peter Adey (2015: 71) puts it: 'An element like air occasionally obtrudes into appearance, but most of the time it is endured and worn, and particularly difficult to abstain from because of its anteriority'. Although air's quality and components for the most part remain intangible to us, in the modern era air has seemed to be an immovable part of our world and our life, even a solid substance – a thing: '[D]uring the period of what we call scientific modernity, we have taken to the air in our tendency to take the air as an object of scientific, technical and philosophical concern' (Connor, 2010: 9). On the one hand, it strikes us as a colourless, odourless, transparent whole and simultaneously as nothingness. On the other hand, we are deeply intermingled with it. Air enters our bodies with every breath we take; it merges with our bodies, and its components are filtered by our vital systems. These components are distributed in different concentrations within our surroundings and our bodies (cf. Choy, 2016). Air itself is both a momentary state and a transforming event. It is simultaneously material and inorganic; natural and anthropogenic; intimate and distributed; freely accessible and political.

Today there is an increasing awareness of the threat that fine dust particles pose to air quality and health. Due to the rise in motorized traffic, mass production and coal-fired power stations, air is bursting with particles that contaminate it, and in many European cities the levels of particulate matter in the air are reaching or exceeding legal limits. Simultaneously, with advancements in sensor technologies and the means of capturing environmental data,¹ public awareness is on the rise. More and more often,

¹ The term 'data' is expressed in the plural following the scientific usage, as explained in the *Oxford English Dictionary (OED)*: 'In Latin data is the plural of datum and, historically and in specialized scientific fields, it is also treated as a plural in English, taking a plural verb, as in 'the data were collected and classified' (quoted by Kitchin, 2014: Note).

therefore, air and its potential contamination are appearing not only on the political agenda of European cities but also within public discourse.

Exploring the sociomaterial production of air

Both globally and locally, environmental monitoring is mainly structured and run by state institutions and their infrastructures. Due to the lack of formal monitoring methods, up to the 1940s the most widespread observations of urban air pollution were simple descriptions of chimney smoke. Even though '[t]hrough the 1920's and 1930's government agencies were designing a range of air pollution monitoring equipment: the jet dust counter, lead candle (SO₂ deposit), filterpaper black smoke monitors and the bubblers' (Brimblecombe, 1998: 16), use of these apparatuses did not become widespread. Like the early measurements of polluted air undertaken by some enthusiasts by the end of the nineteenth century, these efforts were too sporadic and inaccurate to take hold. Only after the London smog of 1952 did the urban monitoring infrastructure of European cities increase substantially (cf. *ibid.*: 16f.). Today the amount of particulate matter in cities' air is primarily measured by permanent and fixed air-monitoring stations. In Germany, for instance (as in many other European countries), the Federal Environment Agency is in charge of the main stations in rural areas. These ground stations are connected to a growing network of European satellites.² In addition to these main stations, within the different cities there are a small number of permanent monitoring stations run by the federal states. These ground stations are installed at specific hubs only; many places remain unmonitored. To increase flexibility regarding where data on air quality is captured, recently so-called 'sentient boxes' have been installed on lampposts and other objects of built infrastructures. They are tested by the cities' governments in cooperation with start-up companies that offer cost-effective, flexible solutions of capturing environmental data through sensors.³

² Copernicus, Atmosphere Monitoring Service: Air quality, <https://atmosphere.copernicus.eu/air-quality>, 28.04.2020.

³ These descriptions are based on my ethnographic research on air monitoring in Hamburg, Germany, and Copenhagen, Denmark (2016-2018).

In addition to these institutional strategies and start-up programs, there are further actors entering the environmental-monitoring stage: a growing number of citizens' grassroots initiatives are also producing environmental data. For the past few years these collaborations have been producing their own nonregulated, open data on air quality and air pollution (see also Gabrys, 2014; Gabrys et al., 2016). They are building their own sensing devices to cover more spots in public spaces; in fact, these sensors now outnumber the permanent stations in Germany.⁴ I have done empirical research on one of these initiatives: the OK Lab Stuttgart, part of Code for Germany, a program of the Open Knowledge Foundation.⁵ The OK Lab comprises local citizens who are interested both in technical tinkering and in the level of air pollution within their city and are building a sensor web across different European cities. The centrepiece of the project's infrastructure is a network of air sensors installed in the environment combined with an open map which shows all of the sensors and their measured data.⁶ Anyone who wants to capture environmental data is welcome to participate in the project. The sensing devices consist of inexpensive hardware components: sensors, a microcontroller, cables, a weatherproof case and open-source software. The codes used are also open source and already prewritten. In addition to the map, the website provides a shopping list with instructions for assembling your own environmental sensing device. People who want to collect data in

⁴ In Germany 365 ground stations are measuring data on air quality on a permanent basis: <https://www.umweltbundesamt.de/en/data/air/air-data>, 28.04.2020. The OK Lab registered more than 1.000 DIY sensing devices, increasing daily: <https://luftdaten.info>, 28.04.2020.

⁵ The OK Lab was founded in Stuttgart, Germany, and reached its peak in 2015. Its initiatives are part of my ethnographic research on different practices of using sensors to produce environmental data in Hamburg and Copenhagen. As part of this research I became a member of the makerspace Attraktor in Hamburg, where people meet who are interested in building their own technical devices that are not connected to the big players of the market. I learned how to tinker with sensors and microcontrollers myself to experience technically mediated environments. The following findings are mainly based on interviews and participatory observations. From January 2016 to December 2018 I conducted about 30 interviews with responsible persons from governmental bodies (in Germany and Denmark) and the European Union, as well as with people from bottom-up initiatives.

⁶ <https://luftdaten.info>, 28.4.2020.

front of their house can build their own movable monitoring station for about US\$30.

Filtering data

Data are ephemeral creatures that threaten to become corrupted, lost, or meaningless if not properly cared for. (Ribes and Jackson, 2013: 147)

In my ethnographic research on these different practices of air monitoring by both state institutions and grassroots initiatives, I observed that data are produced by manifold processes of transmitting, processing and storing. I also found that data production is primarily based not on data collection but on processes that attempt to reduce data quantity and improve data quality. To manage large amounts of data, it is necessary to decrease their number and volume, and constant work must be undertaken to constrict the so-called 'data deluge', for the mere volume of data being collected exceeds institutions' capacity to handle it. Within the data-capturing process itself, the algorithmic filtering techniques are given primary importance.

A filter (from the Latin 'filtrum', used in early alchemists' language) is a 'piece of felt ... for freeing liquids or impure matter' (Hoad, 1996 [1986]: 171) as expressed in the Oxford Dictionary of English Etymology. The two basic functions of the filter device are letting pass and blocking. In this regard, filtering can be characterized as a separating process that subtracts some entities from the total. When incorporated into a certain environment, filters emerge as complex operative systems – as media rather than mere objects – evoking different functionalities like bounding and reducing materials, splitting surfaces, smoothing pressure and temperature gradients, or channelling frequencies. Hence, the consideration of physical, chemical or informational filters is instructive for thinking about the fabricatedness of data. Yet 'filtering' refers not only to the operating mode of sensors in their contact with air but also to the technique of using computerized informational processes to try to make data available. As they are performed by human and nonhuman actors, these processes become social practices.

With the example of air monitoring, the different layers of filtering – physical, chemical, and informational – become quite obvious. Air, as viewed by the

sciences, is the global atmospheric composition of chemical greenhouse gases, reactive gases, ozone and physical aerosols. It does not exist in a 'pure' form, and as a substrate it is neither unified nor solely natural *or* human-made. And in contrast to the ancient view of air as one of the four elements (along with earth, water, and fire), in contemporary science air is understood as a microcosm composed of many elements in different distribution and based on an intense exchange of matter and energy. The particulate matter consists of solid and fluid elements that derive, for example, from stones and the sea as well as from anthropogenic residues (e.g., the black carbon that emerges in industrial burning processes). To gain insight into the density, dispersion, and distribution of particulate matter, researchers filter elements out of the air in different dimensions, by means of complex technologies installed as sensors in the urban realm.

In the ground stations, particulate matter is isolated via physical filters. The filter, with its membrane, is sensitive to the materiality of particulate matter. Material substances are either blocked on the membrane or allowed to pass through. Based on these residues – which have to be dried, as they contain humidity – the density of air pollution is measured and transformed into data. In the temporal and movable do-it-yourself sensing devices used by the OK Lab, the included sensors also serve as filters for measuring particulate matter. In contrast to the physical filters in the ground stations, these filters block light to measure the density and size of particulate matter: reflection on the surface indicates the amount of particulate matter. Both the data-gathering practice itself and the density of air pollution reduce and clean the measured data through informational filters. Specific algorithms such as Kalman filters reduce statistical noise and other inaccuracies by removing redundant or unwanted data. Outliers and scattered data are cleared up, and missing data are added.

These practices of sensing the quality of air within the different spatiotemporal techniques of filtering – in order to capture data – are performed by different actors. In particular, the increasing implementation of sensor technologies in the urban realm enables assemblages of human and nonhuman actors and influences their sociomaterial interplay – based on the notion that computing systems are not only *acting* virtually but are also 'suffused through and through with the constraints of their materiality'

(Blanchette, 2011: 1042). A lot of work needs to be done to hold these complex assemblages of people, places, materials and technologies in place to produce scientific data. The exploration of the different practices of filtering environmental data proves that, as Lisa Gitelman (2013) states, there are no 'raw data'. Data undergo different stages of processing: they are collected, compressed, transmitted, cleaned, visualized, stored and probably also deleted. Thus, they are the result of a series of filtering processes. Unpacking the practices of filtering and thereby focussing on the produced-ness of data makes it clear that data are neither objective (Gitelman, 2013) nor immaterial (Blanchette, 2011). Instead, they arise from a distributed agency of human and nonhuman actors and are produced within the entanglement of environment and technology. The production of data is bound to very different infrastructures and therefore relies on different practices. Data, as elaborated within the concept-metaphor of 'broken data' (Pink et al., 2018), are incomplete, inaccurate and dispersed. A lot of work needs to be done by both humans and nonhumans to produce what is generally thought to be just measured in the sense of recorded. And 'in this context, data ... become a sort of actor, shaping and reshaping the social worlds around them' (Ribes and Jackson, 2013: 148).

This view of data contradicts the contemporary assumption that data are both neutral and omniscient, an assumption that triggers manifold efforts to transform the world into continuous streams of virtually distributed data that are available at any time. Recently, based on the idea of the 'data revolution' (Kitchin, 2014), the 'new promise of omniscience' (Geiselberger and Moorstedt, 2013) and the 'end of coincidence' (Klausnitzer, 2013), the notion of causality has been replaced by a belief in the power of algorithmic correlation of data. The apparent infallibility of large amounts of data has inspired far-reaching predictions of individual and collective futures. The widespread fascination with data's omniscience is driven by the conviction that data are raw, untreated, numeric replications of the world, both keeping it in place available for immediate use at any particular moment in time. The ubiquitous narrative of their power of representation and immediate availability is closely linked to the legitimacy of their incessant capture. But contrary to these common assumptions about captured data, my empirical case study of air measurement indicates that data are neither raw nor

objective; rather, they are produced, and they thus necessarily rest on certain assumptions about the composition of the world. Data are highly reduced through the processes of algorithmic filtering and therefore always have a delay, as the time of the event in the world (e.g., the distribution of particles) is not synchronous with the time of its transformation in data. Filtering, in this regard, becomes a sociotechnical practice that systematically transforms the input into an output – not only by making connections but also by cutting them off.

Data practices becoming political

Bringing these thoughts into dialogue with the different practices of air monitoring indicates that data practices and, concomitantly, the ways of producing scientific facts may become highly political. The aim of the DIY sensing project is to empower people to participate in scientific practice, to encourage a broader understanding of how scientific facts are produced – not only by citizens but also in collaboration with sensing devices, software code and the natural environment. In turn, public authorities of the cities operating the official and fixed monitoring stations have continually contested the data produced by the portable DIY sensors. Their main argument is that the data produced by these devices are vague and not as precise as those produced by the more expensive sensors in the fixed stations. The DIY sensing community dismisses this objection, holding that although the data produced by the individual DIY sensing devices may be a bit less precise, the large number of sensors allows the makers to correct this vagueness⁷ – mainly through various filtering processes.

Another aim of the project is to make data accessible for everyone. Before the emergence of the open data policy (which specifies that data need to be available for public use), the project members saved the data sets of the cities' main stations, which were uploaded on a daily basis before being deleted, and

⁷ As Jennifer Gabrys et al. (2016: 11) observe, '...citizen-gathered data is often "just good enough" to establish patterns of evidence that can mobilise community responses in terms of communication with regulators, requesting follow-up monitoring, making the case for improved regulation and industry accountability, and keeping track of exposures both on an individual and collective level'.

leaked them online. They collected these official data sets and correlated them with their own measurement results. After a few months, they had collected enough data to make a political argument: by correlating the official data with their own, they could show that, in the cities observed, there were several places not yet on the main stations map where legal limits for particulate matter were frequently exceeded.⁸

Rather than the usual data narrative – data as given and immediately available – what I offer here is a notion of data as ‘temperamental and delicate creatures, whose existence and fraternity with one another depend on a complex assemblage of people, instruments and practices dedicated to their production, management, and care’ (Ribes and Jackson 2013: 164). Data may become either publicly accessible or secretly held government knowledge, or they may just go out of date and be forgotten when the next technical advance in distributed storage facilities takes place. Mostly, they remain on hold until they can be useful in political contestations, in commercial cycles or as scientific information. The citizens engaging in the OK Lab experience for themselves that data are produced, are material and are delayed in time. Some data are missing, some are unclear and none are replications of the world; rather, they are translations of existing processes. They are not available instantaneously but need care, and they shape and reshape social worlds.

What I have illustrated here is that within the described sociomaterial assemblages, algorithmic filtering becomes *the* crucial practice in our contemporary world, as massive amounts of data are captured every day, in the environmental monitoring realm and many, many others. The problem we arrive here is that algorithms ‘are opaque in the sense that if one is a recipient of the output of the algorithm ..., rarely does one have any concrete sense of how or why a particular classification has been arrived at from inputs’, as Jenna Burrell (2016: 1) puts it. Data do not just exist; they have to be produced – processed, managed, stored, analysed and utilized. Unless we account for these practices, especially the practice of filtering, the data production

⁸ In this context, Nerea Calvillo (2018) brilliantly elaborates how air becomes political based on a study of the latest conflict in Madrid, where a location change in one official monitoring station caused a drop in the city’s average pollution level.

process remains unclear. Thus, as Nick Seaver (2014: 8) writes, ‘the apparent simplicity and determinism of algorithms in theory does not account for their features in practice’. If we direct more attention to the different ways in which algorithms filter, and thereby produce, the data we work with, it will become obvious that data are neither in constant use nor available immediately. At the moment of data collection, it is often not yet clear to what extent the data may be used or, in fact, whether it will be used at all. As a computer scientist for one of the public authorities states in an interview: ‘We do not yet know how to deal with these massive amounts of data. We try to store as much as possible, but most of the time the captured data are not needed or available in the moment. Our storage facilities are kind of a data cemetery’. Data are being kept on hold until the day they may be of use, and if this never happens, the data may just be deleted or forgotten.

Consequently, these findings force us to reflect on the hopes and presumptions attached to the concept of ‘big data’. The idea behind capturing large amounts of data is that different kinds of data sets may be correlated to produce additional information. But the ability to work well with such massive data sets is still limited because the technologies for correlating the data do not yet reach as far as the theories do. Thus, for now these data are simply stored within distributed and networked infrastructures, whereby their existence is very much tied to their medium of storage. Before the moment they become useful in a productive sense they stay ‘unbound’ within the data warehouses, only loosely related to the information they might reveal in the future. Assuming that information is organised data, what does this new concept of data imply for the notion of information? If data are already organised through the processes of transmitting, processing and storing – all of which involve filtering – then what interpretational act transforms data to information? In line with what has been expressed before, I follow David Beer (2017) in arguing that we need to reveal the technical and material presence of data and the algorithms processing it to understand the emerging worlds we are intermingled with.

references

- Adey, P. (2015) 'Air's affinities: Geopolitics, chemical affect, and the force of the elemental', *Dialogues in Human Geography*, 5(1): 54–75.
- Beer, David (2017) 'The social power of algorithms', *Information, Communication and Society*, 20(1): 1–13.
- Blanchette, J. (2011) 'A material history of bits', *Journal for the American Society for Information and Technology*, 62(6): 1042–1057.
- Brimblecombe, P. (1998) 'History of urban air pollution', in J. Fenger, O. Hertel and F. Palmgren (eds.) *Urban air pollution: European aspects*. Dordrecht: Springer.
- Burrell, J. (2016) 'How the machine "thinks": Understanding opacity in machine learning algorithms', *Big Data and Society*, January–June: 1–12.
- Calvillo, N. (2018) 'Political airs: From monitoring to attuned sensing air pollution', *Social Studies of Science*, 48(3): 372–388.
- Choy, T. (2016) 'Distribution', in *Theorizing the Contemporary, Fieldsights*, January 21. [<https://culanth.org/fieldsights/787-distribution>]
- Connor, S. (2010) *The matter of air: Science and art of the ethereal*. London: Reaktion Books.
- Gabrys, J. (2014) 'Programming environments: Environmentality and citizen sensing in the smart city', *EPD, Society and Space*, 32(1): 30–48.
- Gabrys, J., H. Pritchard and B. Barratt (2016) 'Just good enough data: Figuring data citizenships through air pollution sensing and data stories', *Big Data and Society*, July–December: 1–14.
- Geiselberger, H., and T. Moorstedt (2013) *Big Data: Das neue Versprechen der Allwissenheit*. Berlin: Suhrkamp.
- Gitelman, L. (ed.) (2013) *'Raw data' is an oxymoron*. Cambridge, MA: MIT Press.
- Hoad, T. F. (ed.) (1996 [1986]) *The concise oxford dictionary of english etymology*. Oxford: Oxford University Press.
- Kitchin, R. (2014) *The data revolution: Big data, open data, data infrastructures and their consequences*. London: Sage.

Klausnitzer, R. (2013) *Das Ende des Zufalls: Wie Big Data uns und unser Leben vorhersagbar macht*. Salzburg: Ecowin.

Latour, B. (2006) 'Air', in C.A. Jones (ed.) *Sensorium: Embodied experience, technology, and contemporary art*. Cambridge, MA.

Pink, S., M. Ruckenstein, R. Willim and M. Duque (2018) 'Broken data: Conceptualising data in an emerging world', *Big Data and Society*, January–June: 1–13.

Ribes, D., and S. Jackson (2013) 'Data bites man: The work of sustaining a long-term study', in L. Gitelman (ed.) *'Raw data' is an oxymoron*. Cambridge, MA: MIT Press.

Seaver, N. (2014) 'Knowing algorithms', *Media in Transition* 8. [<https://static1.squarespace.com/static/55eb004ee4b0518639d59d9b/t/55ece1bfe4b030b2e8302e1e/1441587647177/seaverMIT8.pdf>]



Figure 1: My tinkering kit

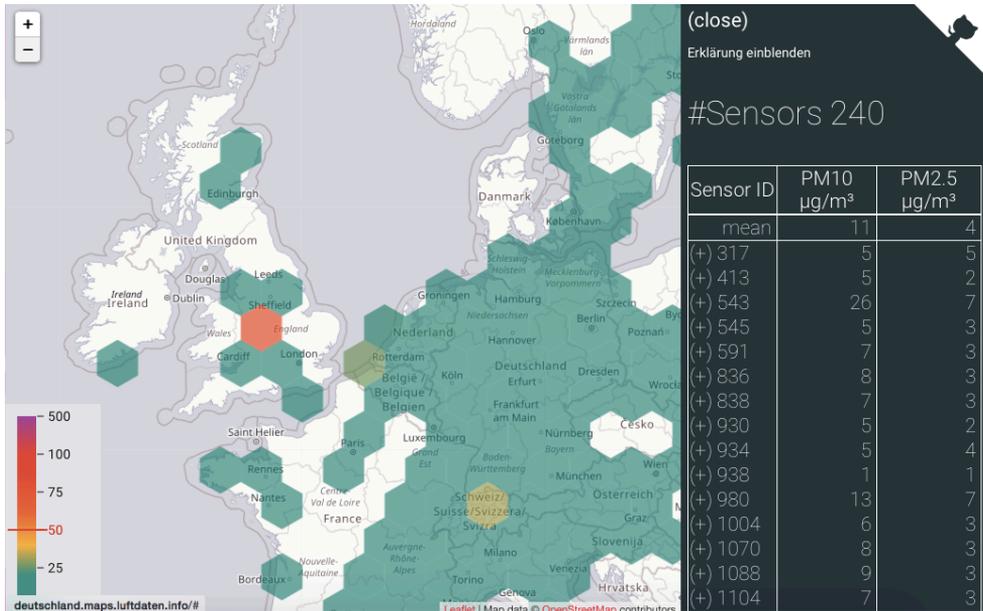


Figure 2: Sensor map with cleared-up data (15.06.2018 luftdaten.info)

Station	Habichtstraße						
Komponente	CO	NO	NO ₂	Benzol	Toluol	NO (4m)	NO ₂ (4m)
Einheit	mg/m ³	µg/m ³					
Messzeit	Stundenwerte						
30.03.2017 18:00	0.8	127	120	81	101		
30.03.2017 17:00	0.63	163	123	2.9	106	104	
30.03.2017 16:00	0.6	194	118	1.7	5.6	119	102
30.03.2017 15:00	0.6	172	100	1.3	2.3	115	90
30.03.2017 14:00	0.71	155	96	1.5	6.2	119	91
30.03.2017 13:00	0.5	116	82	1.1	4.5	91	79
30.03.2017 12:00						93	79
30.03.2017 11:00						140	85
30.03.2017 10:00						173	96
30.03.2017 09:00	0.99	275	121	2.4	8.1	249	118
30.03.2017 08:00	1.07	281	123	2.9	8.5	257	120
30.03.2017 07:00	0.92	224	108			203	105
30.03.2017 06:00	0.44	91	71	1.3	4.3	79	69
30.03.2017 05:00	0.22	34	41	0.5	1.4	28	40
30.03.2017 04:00	0.1	16	31	0.7	3	4	11
30.03.2017 03:00	0.21	12	33	0.7	3	6	9
30.03.2017 02:00	0.24	5	21	0.5	1.4	2	20
30.03.2017 01:00	0.22	10	31	0.5	1.4	7	29
30.03.2017 00:00	0.25	19	36	0.6	1.5	12	32

Figure 3: CSV data uncleaned (Monitoring station Habichtstraße 30.03.2017, 6pm, luft.hamburg.de)

Einheit	CO ⁽¹⁾	NO ⁽¹⁾	NO ₂ ⁽¹⁾	Benzol ⁽¹⁾	Toluol ⁽¹⁾	NO (4m) ⁽¹⁾	NO ₂ (4m) ⁽¹⁾
Messzeit	mg/m ³	µg/m ³	µg/m ³	µg/m ³	µg/m ³	µg/m ³	µg/m ³
	1h	1h	1h	1h	1h	1h	1h
30.03.2017 18:00	0.8	127	120			81	101
30.03.2017 17:00	0.63	163	123	2	9	106	104
30.03.2017 16:00	0.6	194	118	1.7	5.6	119	102
30.03.2017 15:00	0.6	172	100	1.3	2.3	115	90
30.03.2017 14:00	0.71	155	96	1.5	6.2	119	91
30.03.2017 13:00	0.5	116	82	1.1	4.5	91	79
30.03.2017 12:00						93	79
30.03.2017 11:00						140	85
30.03.2017 10:00						173	96
30.03.2017 09:00	0.99	275	121	2.4	8.1	249	118
30.03.2017 08:00	1.07	281	123	2.9	8.5	257	120
30.03.2017 07:00	0.92	224	108			203	105
30.03.2017 06:00	0.44	91	71	1.3	4.3	79	69
30.03.2017 05:00	0.22	34	41	0.5	1.4	28	40
30.03.2017 04:00	0.1	16	31	0.7	3.4	11	29
30.03.2017 03:00	0.21	12	33	0.7	3.6	9	31
30.03.2017 02:00	0.24	5	21	0.5	1.4	2	20
30.03.2017 01:00	0.22	10	31	0.5	1.4	7	29
30.03.2017 00:00	0.25	19	36	0.6	1.5	12	32

Figure 4: Same data after being filtered: outliers have been reduced and missing data added (Monitoring station Habichtstraße 30.03.2017, 6pm, luft.hamburg.de)

the author

Vanessa Weber is researcher in the project „Digital Curation in the City“ (funded by the German Research Foundation) at City Science Lab of HafenCity University, a cooperation with MIT. Her research interests lie in the field of cultural sociology and related disciplines such as media studies and science and technology studies (STS). In particular, she is interested in algorithmic cultures and the materiality of the social as well as their aesthetic, sensory, and affective dimensions.

Email: vanessa.weber@hcu-hamburg.de