ESSnet Big Data II

Grant Agreement Number: 847375-2018-NL-BIGDATA

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata
https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en

Workpackage L

Preparing Smart Statistics

Deliverable L3: Description of the findings regarding Task 3, Smart Devices

Final version, 30th October 2019

Prepared by:

Olav ten Bosch (CBS, the Netherlands)
Sónia Quaresma (INE, Portugal)
Massimo De Cubellis (ISTAT, Italy)

Workpackage Leader:

Natalie Rosenski (Destatis, Germany)
Mail address: natalie.rosenski@destatis.de
Telephone: +49 611 754284

Table of contents

1 Executive summary

This deliverable describes the results of the ESSnet Big Data II, Workpackage L (WPL), task 3 on *smart devices*.

Data generated by smart devices can be of special interest to official statistics. Such devices, generally connected to other devices or networks via wireless protocols and operating to some extent interactively and autonomously, may generate huge amounts of digital observations that are a promising input for both traditional as well as experimental official statistics. Examples are smartphones, smart cameras, human wearables, medical devices, tracking apps/devices, smart home devices and security equipment. This deliverable contains a study on the use of smart device data from the viewpoint of official statistics.

One key prerequisite is access to data. The report contains some reflections on possible mechanisms to actually access smart device data such as: direct access, pushing computation out, access via manufacturers' portals and using the concept of citizen science. In all cases it is necessary to build partnerships. Partnerships with other entities already focusing on the application of smart sensors in society and providing similar gatherings of sensor data. For this, it is good to keep in mind the difference between public sector data, private sector data and community data as partnerships in the first and latter case are probably the better starting point.

When using data from smart devices for official statistics, attention has to be paid to the representativeness of the data. Both the coverage of smart devices with respect to the population of interest as well as a possible bias in the data because device users may differ considerably from non-device users should be addressed. Such representativity issues vary considerably per device type and per user community and should be studied on a case by case basis.

The core element of this study is a long list of smart devices organized in the categories smartphones, smart home devices, smart devices for health and fitness, smart devices for mobility, smart devices for travel and other smart devices. Every item of this list contains a description, some thoughts on their data use for official statistics and some reflections on popular categories or brands. From this list we conclude that smartphones - the most widely spread type of smart device – with their ever growing number of integrated sensors is certainly a candidate for further experiments. Data from smart thermostats can be promising to gain insights in energy consumption patterns of households. Networks of intelligent sensors – as used for citizen science air quality projects or traffic measurements – are promising as well, as a community approach to gain real time measurements of statistical phenomena. Smart devices for healthcare and fitness could provide data that are not easy to measure otherwise, but are less straightforward to start with, due to ethical and privacy considerations. Smart devices for public and private transportation as used in smart cities can help produce indicators both in the context of mobility and in the social context.

This study is meant to be preparatory for a possible next project in the area of Trusted Smart Statistics. However, we realize that the long list in this document is just a 2019 snapshot in a fast changing society getting more digitized every year. We cannot possibly foresee how the smart device landscape will evolve. Keep in mind that this is not only depending on technical developments, it is also a peoples' choice how to live their digital life, which is not always easy to predict. The last chapter of this study therefore contains three conceptual use case decriptions that are meant to be inspirational for a possible follow-up. They describe the concept of citizen science networks of smart devices collecting nearly real-time indicators on certain phenomena, the use of smartphones and integrated sensors in combination with machine learning software to derive behavioural patterns and the use of smart travel cards to understand travel patterns and other phenomena of mobility.

One conclusion from this study is that the categories of smart devices are very different in terms of functionality, use and data opportunities. Nevertheless, although it was not easy to find concrete use cases that would deliver an added value for official statistics on the short term, we hope that this study inspires and helps trusted official statistics getting even smarter.

2 Introduction

This deliverable describes the results of the ESSnet Big Data II, Workpackage L (WPL), task 3 on *smart devices*. Smart devices are expected to be of special interest for future official statistics, especially the future variant known as Trusted Smart Statistics. The use of these devices producing digital footprints is growing and the data they generate is therefore a promising input for both traditional as well as experimental official statistics. Examples are intelligent sensors, human wearables, medical devices, tracking apps/devices, smart home devices and security equipment. These devices are very different in their nature, not only in terms of the functionality and use, but also in terms of communication and physical implementation. All these aspects have an implication on their potential use for official statistics.

The aim of this task is to give an overview on smart devices and - wherever useful and feasible - citizen science data from the viewpoint of official statistics and to explore some practical use cases. It is meant to be preparatory for future projects on Trusted Smart Statistics that will be able to go deeper into the most promising use cases identified in this task.

2.1 Scope and definition

Generally speaking, data from any smart device or citizen science project could be of interest for producing official statistics as a direct or auxiliary data source. Therefore, we take a very broad approach, first putting together a long list of smart devices per domain of use, then reviewing them from the perspective of data value and data accessibility from the official statistics perspective.

Regarding the definition of a smart device, at the kickoff meeting of this workpackage it was agreed to take the definition from wikipedia[1] as a starting point:

*A **smart device** is an electronic device, generally connected to other devices or networks via different wireless protocols such as Bluetooth, NFC, Wi-Fi, LiFi, 3G, etc., that can operate to some extent interactively and autonomously.*

Two key elements in this definition are the observation that the smart device has to be *interactive* and *autonomous* to some extent. Interactive meaning that users can somehow influence the behaviour of the device, autonomous meaning that the device can operate for a longer time without human intervention to do its task.

The wireless protocols mentioned in the wikipedia definition are to be taken as an example as new wireless protocols are added continuously[2] these days, some dedicated for specific use, such as long distance or low power communication, others developed by cooperating smart device producers targeted at a specific area of use, for example home devices.

One could ask how smart devices relate to the popular concept of the Internet of things (IoT)[3], known as *the extension of the internet connectivity into physical devices and everyday objects*. One could argue that the concepts are closely related and in some cases the terms are used to denote the same concepts. However, in this study we will use the following view: a smart device is part of the

[1] https://en.wikipedia.org/wiki/Smart_device as of 2019 June 14th
[2] https://www.rs-online.com/designspark/eleven-internet-of-things-iot-protocols-you-need-to-know-about
[3] https://en.wikipedia.org/wiki/Internet_of_things as of 2019 June 14th

IoT because of it connectedness, but not every IoT object is a smart device because it might lack interactiveness and autonomy.

2.2 Eurostat reference architecture and types of data access

At the DGINS conference 2018 Eurostat presented a reference architecture for TSS[4]. One of the core concepts is to replace – or at least add to – the traditional approach of "*pulling data in*" – from data sources to statistical offices – to "*push computation out*" – from statistical offices backwards towards the data source. Putting freely, instead of processing the – maybe massive – data streams on the microdata level, the NSI agrees with the data provider on an algorithm to calculate aggregates at just the right level for official statistics.

In the context of this task on smart devices the concept of pushing the computation out is very interesting. We see some variants. The computation could not only be done by the device manufacturer / data owner based on an algorithm specified by the NSI, there are a few variants possible, such as:

- The NSI interprets data from a publicly accessible data portal where the data owner publishes the data in an *aggregated* form. An example is the interpretation of use of public space from the heatmap of Strava[5] (running and cycling) app. Research has shown that, although difficult, this is possible. In principle this could hold for data from any smart device if it is published in an aggregated form on a manufacturer's portal.

- Smart device owners that upload their data to a *citizen science* portal where it lives as open data ready for processing by anyone, including statistical offices. Examples are air quality measurements by citizens, published on data portals, such as openSenseMap[6]. Also citizens observing ADS-B[7] signals from airplanes adding these data to open data portals fall into this category. There might be more.

We conclude that the main concept of pushing the computation out to smart device data holders can indeed be a valuable concept, but also that we should be open to variants such as using data from manufacturers portals and citizen science projects. To complete the picture we note here that it is theoretically also possible that statistical offices *directly* subscribe to data streams from smart devices.

Thus, we distinguish between four types of forseeable data access to smart device data:

- ***Direct access:*** Although not the most common case, it might in principle be possible that the NSI receives (or intercepts) data from the smart devices directly. For example imagine smart traffic lights sending their data not only to the traffic regulation authorities, but also directly to the statistical office. This could help create real-time traffic density statistics.

[4] Towards a Reference Architecture for Trusted Smart Statistics, Ricciato et al., DGINS 2018
[5] https://www.strava.com/
[6] https://www.opensensemap.org/
[7] Automatic dependent surveillance—broadcast, see https://en.wikipedia.org/wiki/Automatic_dependent_surveillance_%E2%80%93_broadcast

- ***Pushing computation out:*** An obvious example here is the case where mobile network operators run aggregation algorithms designed together with data scientists from the official statistics community on their computing infrastructure, sending the aggregated (anonymized) results to the statistical office for further processing.

- ***Manufacturers' portals:*** In this case one or more device manufacturers publish (usually aggregated) data on the use or results of the smart device they manage. Even such highly aggregated data can be useful for tracking phenomena, especially if it is monitored frequently to see changing data patterns. An example is the reverse engineering of sports activities on regional scale from the aggregated heatmaps of Strava, but producers of other smart device categories might do something alike.

- ***Citizen science***: As explained above in certain domains, citizens tend to do measurements themselves using smart sensors that they connect to one or more citizen science portals that publish the data, usually as open data. A well-known example is the *Luftdaten* network of air quality measurement sensors[8], but other citizen science projects, such as for traffic measurements, can also be found.

2.3 Public sector data, private sector data and community data

On 22 January 2019, negotiators from the European Parliament, the Council and the Commission reached an agreement on the Directive on Open Data and the re-use of public sector information[9]. The agreement extends the scope of the Directive on access to and re-use of public sector information (so-called PSI directive) to also cover public undertakings operating in some sectors (water, energy, transport and postal services sectors) and those acting as public service operators for the transport of passengers by rail, road, air or maritime carriers. Furthermore, according to the directive, dynamic data (such as data generated by sensors) will be available for re-use immediately after collection.

The extension of this directive is positive for the use of data from smart devices by official statistics. Although the actual implications in practise will become clear within a few years, we can already imagine that the opportunities for NSIs having immediate access to smart device data produced by or for the public are numerous. Imagine immediate access to all smart data that is generated in the transport systems of municipalities for processing statistical indicators on mobility, transport, energy use and maybe also time use and other behavioural patterns. Therefore, we feel it is useful to keep in mind whether a smart device category is essentially delivering public sector data.

Moreover, we feel that the categorisation of types of data into public sector data and private sector data is not complete. Especially with the rise of citizen science projects using smart devices to do some measurements and publishing their data as open data, we feel that we should add another category: community data. Probably there are even more variants of data access methods and types of data possible, but before philosophizing about that, it is probably better to touch the ground and dive into examples of smart devices as we do in the next chapter.

[8] See chapter 3.2, Smoke / CO sensors / Air quality systems

[9] http://europa.eu/rapid/press-release_IP-19-525_en.htm

For all types of data, building partnerships is crucial. Partnerships with other entities involved in data-collection by means of smart sensors, such as – but not limited to – companies and governmental bodies working on smart cities, environmental agencies, public entities of transportation. This holds for all types of data, however since partnerships are probably easier to build with public entities than private ones, we suggest starting there.

2.4 Representativity

When we study the use of data from smart devices for producing, improving or supplementing official statistics – as we do here - one may ask the question how *representative* these data would be for calculating statistical estimates on a population. Some smart devices may be used mainly by a group of front-runners, which makes it difficult to make inferences on a larger population. Also, the people using the devices might differ considerable from those not using them or not willing to share their data, leading to a possible bias in our estimates. Both the *coverage* of smart devices, with respect to the population of interest, as well as the *bias* in the data, as described above, should be addressed when using smart device data for official statistics.

One observation to make here is that data from smart devices do not only say something about the behaviour or activities of the device owners. In many cases they also say something about the environment or circumstances in which the device is being used. Thinking from the example of a sports device, the data does not only say something about the sporting activities and health conditions of the device owner, they also say something about the use of public spaces for a particular type of activity at a certain point in time. The same applies to smart devices for travel and transport. They do not only comprise data on the behaviour of the traveller, they also provide data on the use of the travel system as a whole. Since official statistics is about describing society and economy as a whole, we are – in addition to data on the device owner - also particularly interested in this second type of use case, where smart devices make observations on their context. Of course, also in these cases one should pay attention to representativity issues. For example users of smart sport devices might prefer to do their exercises in a more urban environment than non-device users. Similar issues might apply to users of home, transport or health devices. From a statistical perspective this is not a showstopper at all. If such structural biases are known, we can account for it in our statistical estimates.

Generally speaking, the representativity issues to be addressed when using data from smart devices will largely vary per device type, per community and probably also over time. In one case statistics on the use of a particular type of smart device might be a valid use case as such, in other cases we might use the device data to gain insights into phenomena to which the smart device indirectly interacts. Any of such use cases could be valuable to us. Just like a gold digger, we do not know if the use case is there, unless we start digging. That is exactly what we will do in the next chapter: we start digging around for expected or unexpected gold, and we do so by exploring different types of smart devices freely.

3 Smart devices: a long list

3.1 Smartphones

The most widely spread smart device, which is used by an immense number of people around the globe, is without any doubt the smartphone. These days, a smartphone is a powerful computer that may run hundreds of system and user-installed apps, each performing a specific function that may require interaction and may also operate autonomously where needed. Multiple wireless connection protocols are supported, typically WiFi, bluetooth, NFC and 3G, 4G or even 5G. Depending on the type and brand, a smartphone may contain multiple advanced sensors, such as an accelerometer, gyroscope, magnetometer (compass), GPS, microphone, camera(s), touch and fingerprint sensors, which make it a powerful observation device in itself. It generates lots of (privacy-sensitive) data, but it all depends on the apps being installed and their specific use. Ideally, we would present a list of apps here that are most promising from the perspective of official statistics, if the data – micro or aggregated – would be accessible, but the list would be huge. So for the moment we just state that app providers in general might be a candidate to "push some computation out". Also, we mention the possibility for statistical offices to provide their own app (which is already a fact for replacing traditional surveys, such as for example the time use survey).

3.2 Smart home devices

Smart home equipment is getting cheaper and more advanced these days. Where one had to buy expensive "domotics"[10] before, now consumers can buy lots of smaller and cheaper smart home devices with various functions that may connect to each other using WiFi or more specific indoor communication protocols such as Bluetooth Low Energy, Zigbee etc. Moreover, smart home devices – even from different manufacturers – may be programmed in a chain of logic, using a customer-friendly programming enviroment, such as IFTTT[11] (if this than that).

Below we present a non-exhaustive list of categories of smart home devices[12], with some early reflections or expectations on its use for official statistics.

Wireless speakers: One of the device categories that recently improved their smartness are wireless speaker systems. Well known examples are Amazon Echo, Google Play, Apple HomePod and Sonos play. They connect via WiFi or use a dedicated networking protocol for improved performance. These days, the more advanced wireless speakers have a voice interface which makes it possible not only to activate sound-related services, but also to interact with your digital assistant. Examples are Alexa from Amazon, Google's Assistant and Siri from Apple. It is interesting to see that this type of smart devices quickly becomes the voice interface to a wider set of personal and household services, such as consulting your agenda, the news, travel information or communicating with other smart home devices. Roughly speaking, the basic function of music management might not be too much of interest for official statistics, the wider use of controlling other smart devices and a voice controlled assistant function are very interesting.

[10] Domotics refers to the a home automation system typically connecting controlled devices to a central hub; https://en.wikipedia.org/wiki/Home_automation

[11] https://ifttt.com/

[12] Inspired by http://iotlineup.com

Smart TVs: There are literally too many brands to name here. These devices generally connect via WiFi and can be used to install the same kind of media apps that can be used on a smartphone or a tablet. The data generated by these apps might be somehow interesting for official statistics, however with the growing popularity of the smartphone and its ability to stream media – video as well as audio – one can expect that the use of these apps will be less dominant compared to those on smartphones. Also, a smartphone has many more sensors built in and is usually more personally connected to one person and thus probably more interesting at this stage for official statistics purposes.

Smart thermostats: In this category, we find many specific smart themostats from energy producers, but also a few generic players, such as the Google Nest or products from Ecobee4, Honeywell etc. They typically detect patterns in energy use to build a smart schedule for temperature control. From a data point of view, these patterns say something, fairly reliable and quantitative, about energy use of households and can therefore be of interest for official statistics.

Kitchen appliances: Examples are smart coffemakers connected to the WiFi to be monitored and operated from a distance, ovens that can be operated via an app or provide a display for easy recipy consultation and smart refrigerators[13] that can detect the items stored and keep track of expiry and usage (also known as internet refrigerator). With a bit of imagination, maybe in some years smart refrigerators will even communicate with supermarket delivery services to autonomously top up essential food products. If this happens, kitchen appliances could provide valuable insights into food consumption patterns. Although not immediately feasible, in the long run these smart devices might be a valuable source of data for official statistics.

Smoke / CO sensors / Air quality systems: Especially for modern houses with good isolation, the monitoring of indoor air quality is important. Low cost indoor air quality monitoring devices are becoming mainstream. To be valuable, they increasingly interact with other home systems to collectively manage a healthy environment. Fire prevention systems such as CO and smoke sensors are also getting smarter, they do not only implement the basic function of alarming in case of fire / pollution, they continuously deliver their data to other home controllers. This data is certainly interesting, as it might supplement outdoor measurements to give the full spectrum of air quality conditions for people.

We want to give a special not to smart devices for measuring outdoor air quality. There are several citizen science projects using smart devices for air quality measurement that deliver their data in real time to a cloud of community data. A well-known example of such a citizen science project is the *Luftdaten* project[14], which originally started in Germany and is now spreading around Europe. Depending on the sensors installed, the smart device produces time series of measurements on variables such as pm10, pm2.5, NOx, humidity, temperature etc. The owner of the smart device decides where to deliver the data. An example of a data portal showing an impressive number of measurements internationally on various variables is the opensensemap[15] portal. One may wonder whether such a setup - smart devices measuring something connected to a central processing unit, processing and publishing in real time - could be used in Trusted Smart Statistics. We come back to that in chapters 4 and 5.

[13] https://www.techopedia.com/definition/15684/smart-refrigerator
[14] https://luftdaten.info/en/home-en/
[15] https://opensensemap.org/

Garden equipment: Examples are smart water management systems, smart garden lighting systems and garage door automation. Their smartness might rely on sensors, for example measuring humidity. This type of sensor data is available from other sources as well (open data from meterological institutes) and do not provide much insight in human activities. So at this moment, we feel it is not particularly interesting for official statistics.

Consumer security cameras / Alarm control units / Tracking devices / Smart locks: There are a lot of home devices for security that have an increasing number of smart features. Cameras can be instructed to alarm when people enter that are not one of the recognized residents or pets. A smart doorbell has motion detection and sends high definition video to your phone immediately offering two way communication possibilities with the outsiders. Small tracking devices make it possible to find anything if lost. Smart locks provide wireless (un)locking facilities from authorized smartphones. It is to be expected that smart devices for security will develop further and will provide additional security services. However, from an official statistics perspective we do not see much data interest at this moment.

Robotic pets: A well known example of a WiFi connected robotic pet is the Aibo by Sony. The first generation was launched in 1999, it was refined several times up to the fourth generation launched in 2018. Although it can be seen as a toy (a very expensive one) it is also known as a companion robot for adults. There is a software development environment for people that want to program it. With a little bit of fantasy one could imagine that (cheaper) robotic pets could provide a (voice controlled) assistance function as was described above for wireless speakers. Clearly there is no short term data value for official statistics, this is something to keep an eye on for the long term as robots in houses could generate interesting sensor and performance data on daily life.

3.3 Smart devices for health and fitness

Smart devices for health have been widely introduced in the market not only enabling medical organizations to provide healthcare with reduced costs and improve patient treatment, but also to the citizens who can now monitor themselves and decide when and how to look for medical support. In this field, smart devices are already changing the landscape and will enable the citizens more and more to make informed choices as they can buy several small and cheap portable devices to keep track of their own health or of loved ones.

The possibility of preventive analysis, the use of connected monitoring wearables and health apps on any smartphone also makes it possible for the general public to monitor their lifestyle and fitness performance. Therefore, more and more people are using smart devices to track their own physical condition and to help improve it. Although these categories of smart devices target a slightly different area, we describe them in the same section.

The smart devices can vary widely and are also different depending on the targeted user, i.e. a medical organization or a citizen without medical skill or knowledge. However, most of them can become important sources for health, life conditions and life satisfaction statistics and even to detect areas where new surveys are required. The fact that devices/apps are easier or harder to use for medical purposes or increase the patient’s comfort, will potentially have high impact in the device/app being broadly used/adopted. In this way, we must pay more attention to it in the future. For this reason, the explanation of the device is sometimes more comprehensive. We are also aware that this data will target only subpopulations, as almost every data that comes from sensors/apps.

However, health statistics may target the same subpopulations. In Europe, due to the aging population problem, the health and life condition statistics are becoming more important and receiving more attention from governments and society, which justifies our effort in trying to get acquainted with the smart devices for health and fitness market. None of these devices/apps are very widespread yet as the technologies presented are often cutting edge and it will take time for them to spread. On the other hand, it could be the perfect timing to develop partnerships with hospitals/doctors providing them with our statistical skills in exchange for their anonymized raw data.

Finally, the devices presentation is divided into two categories: 1. to facilitate diagnosis/monitoring and 2. to facilitate treatment. The reason for this is that the ones to facilitate diagnosis will probably target larger populations and be less intrusive in terms of personal data if the data is pre-aggregated first for example. Of course, in both cases the data is much more sensitive in nature than data collected by other wearables, thus requiring anonimization, aggregation and built-in security to avoid individual identification. Still, someone requiring medical treatment for a particular condition will potentially be more vulnerable and less willing to share the data than the larger population monitoring their activities. In the latter case, by aggregating information some level of protection is provided. Thus they are presented first, as they will probably show higher potential in the short term for statistical data discovery and exploration.

1. ***To facilitate diagnosis/monitoring***:

The smart devices introduced to facilitate diagnosis are mostly Biosensors that transmit medical or fitness information over a wireless network to mobile and web applications. The Biosensors collect data continuously during the person's physical activities and about the activities they perform in their everyday life. This ability to keep track of the person's condition in real time makes them very useful for health management, allowing alerts to the self and/or to the doctors/hospitals to be issued immediately should a critical situation arise.

Devices that measure glucose levels: Diabetes affects over 400 million people. To control the level of their blood sugar, patients must test it several times a day, usually by pricking their finger with a lancet. This is uncomfortable and painful, which may result in less frequent testing and therefore a poorer control of blood sugar levels. Several portable/wearable devices offer a needle-free solution. Some of them are not invasive like Sugarwatch by EPSBio that takes its measurements through the skin. Some use disposable parts like FreeStyle Libre's small sensor patch that is placed on the arm and can be worn for up to 14 days. Others are long lasting as Noviosense glucose sensor with a flexible metal coil of just 2cm in length containing nanosensors inside and covered by a protective layer of soft hydrogel. This sensor is placed under the lower eyelid, from where it can wirelessly send glucose measurements directly to a smartphone. Most sensors for measuring glucose levels rely on a combination of ultrasonic, electromagnetic and thermal waves like GlucoTrack and are capable of sending the data directly to a smartphone or at least to some form of online dashboard.

Devices that measure arterial pressure: Blood pressure is a critical vital sign and patients, particularly those with cardiovascular conditions, are often directed to take readings on a regular basis. There are already a lot of measuring devices, mostly relying on cuffs or some kind of prop that sends the data to the smartphone over Bluetooth. Some of the portable/wearable devices like QArdioArm or Omron Cuffs will probably not be so interesting in the future due to the recent technology that through an optical sensor will allow users to measure their blood pressure, without using an ancillary device. There are already initiatives with universities or other medical institutions to further research in the field of vital signs monitoring such as My BP Lab. Smartphones and watches, also because of the extensive number of sensors they include, may be used for this monitoring. Please see section 3.1 in this regard.

Devices that measure heart rate: A common method for measuring the heart rate is photoplethysmography (PPG). It works by shining a light onto a patch of skin and measuring the reflection with a light sensor. The way light is scattered, and thus the amount of light reflected, varies depending upon blood flow and therefore changes with the pulse. Algorithms detecting these changes are able to estimate the heart rate based on the fluctuation of light being reflected. The modern smartphones are able to use their flash as the light source and the camera as the sensor. An app can therefore calculate a measurement comparable to a medically-approved professional device.

However, apps using the same hardware may provide different results. First, because there are two forms of using PPG: contact and non-contact. When using the former, you place your finger directly over the camera, while non-contact apps require you to simply point your camera at your face. And second, because of the way in which the algorithm is implemented. Most popular choices are Instant Heart Rate, Runtastic Heart Rate Monitor, Cardiio or Sleeptracker 24/7. Some popular apps like Qardio, although tracking more metrics like the blood pressure, require an extra device as the CardioArm to work. Smart watches have an advantage over the phones being in constant contact with the skin while worn, so are able to monitor the heart rate without the need to perform a specific action. The main drawback with tracking the heart rate using a smart watch is that PPG monitoring struggles to cope with movement. That is the reason why runners or people engaging in sports usually measure the heart rate using arm, wrist or chest bands. The most well known chest straps like Polar H10 or Garmin HRM (Run, Swim or Triathlon) are found to be accurate (over 99%) across all exercise conditions, when compared to a professional-grade electrocardiogram (ECG). The type of data delivered by all these devices has very similar characteristics.

Devices that measure oxygen level: The oxygen level on the blood or more accurately the oxygen saturation measuring is known as pulse oximetry and is a non-invasive monitoring method. A sensor device is placed on a thin part of the patient's body, usually a fingertip or earlobe, or in the case of an infant, across a foot. The device passes two wavelengths of light through the body part to a photodetector. It measures the changing absorbance at each of the wavelengths, allowing it to determine the absorbance due to the pulsing arterial blood alone.

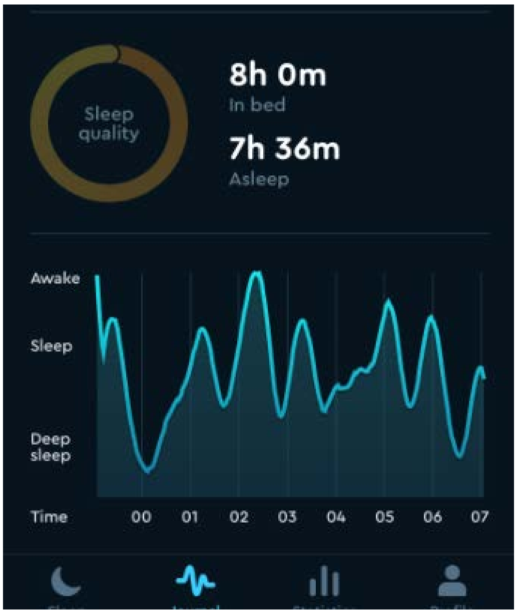
Some oximeters are most commonly used in a medical environment and others in fitness and lifestyle monitoring situations. Basically they work in the same way and use a finger wearable like iHealth Pulse Oximeter or a wrist wearable like Vincense or Visi Mobile. The Visi Mobile for example is used in hospital wards to collect data for calculating Early Warning Scores (EWS). EWS are physiological track-and-trigger systems, which use a multiparameter or aggregate weighted scoring system that assists in detecting physiological changes and thereby identify patients at risk for further deterioration. Continuous monitoring of vital signs could be a useful tool to detect clinical deterioration in an earlier phase, avoiding the interval measurements that may not capture the changes of vital signs. The devices mentioned are most used in medical related environments and all monitor other data like heart rate, blood pressure or even ECG and the type of data collected across the devices is analogous.

The oximeters are also important to check whether breathing stops during sleep and to measure a person's ability to handle intensive physical activities. Two of the most well known are developed by Fitbit and Garmin. Garmin focus on measuring oxygen saturation and relating the oximeter readings to the changes in the subject elevation through what they call the Pulse Ox Acclimation. Their device the Fenix 5X Plus watch targets strenuous sports like hiking, alpine sports and going on big expeditions. Fitbit wants to tackle sleep apnea and both its Ionic and Versa smartwatches include a light-based SpO2 sensor, which is a pulse oximeter that measures blood oxygen levels.

Devices that monitor sleep: Sleep apnea, introduced with oximeters, is not the only major sleep disorder. To assert the quality of sleep, not only its duration, but also the sleep stages need to be tracked. Several smart watches as well as fitbits already monitor this information, not only tracking sleep, but providing hints on how to improve it and even using alarms, sometimes silent, to wake the person gently at the optimum point of their sleep. Some wearables like Neuroon go a little bit further when monitoring the switch from mono-phasic to poly-phasic sleep. This is achieved through measuring electroencephalogram (EEG) and electrooculogram (EOG), besides the most common pulse, motion and temperature data.

Another kind of smart devices devoted to sleep monitoring are the ones designed for babies to detect heart rate, oxygen levels, skin temperature and sleep quality. The wearable devices are usually worn on the ankle of the baby with straps. The main difference here is not on the sensors, but on the location and the market. Also, it is probably more sensitive data that is not easily available.

Both devices provide data in a graphical format that must be converted to data points to become useful.



8h 0m
In bed

7h 36m
Asleep

Devices that measure brain activity: For several years, several states of wellbeing, creativity or intense focus have been studied through the brain waves of the people exhibiting these behaviors. Five different types of brain waves can be detected through EEG sensors in a headband. Gamma Brainwaves (Frequency: 32 – 100 Hz) are associated with states of heightened perception, learning and problem-solving tasks, while Beta Brainwaves (Frequency: 13-32 Hz) are found in alert individuals, that exhibit normal alert consciousness or active thinking. Associated with relaxation are the Alpha Brainwaves (Frequency: 8-13 Hz) present in physically and mentally relaxed persons or even at lower frequencies the Theta Brainwaves (Frequency: 4-8 Hz) displayed by subjects in moments of creativity, insight, dreams and reduced consciousness. For sleep monitoring, the Delta Brainwaves (Frequency: 0.5-4 Hz) are the most important ones as they indicate restorative sleep in a dreamless state, where healing and rejuvenation are stimulated.

The Delta Brainwaves as well as other variables monitor sleep. However, due to the recent attention meditation, yoga and other focusing techniques have been receiving to enhance well being, some devices are available on the market to measure the brain activity of awoken subjects to assist them in their meditation and to give them feedback in their neurotraining sessions. Some examples are the Muse headbands and the Unicorn Brain Interface. In both cases, the brain activity can be visualized and explored through apps of the smartphone. Muse's brain-sensing technology has been widely used for brain research projects in hundreds of hospitals and universities worldwide, including NASA, The Mayo Clinic, UCL and MIT, which leads us to believe that anonymized datasets may exist. However, the interest of these brain usage maps seems to be more for specific medical situations and not for statistics in general.

Devices that measure food intake: Food has a key role to maintain the overall wellbeing of the organism. While on the one hand it can improve quality of life, increase life expectancy and help prevent serious diseases, on the other hand, unhealthy food can be one of the drivers for serious health problems that can lead to the chronic diseases like diabetes, obesity and heart disease. These are the facts that make it potentially interesting for statistical production, albeit its current novelty status in the market.

Bearing in mind that each metabolism is unique and needs such (unique) treatment, many intelligent systems have been developed to help people achieve their personal nutrition goals. There are various devices on the market that collect data on eating habits and food consumption through sensors and that perform complex algorithms and machine learning to recommend health and fitness programs. Most devices are kitchen appliances, such as HAPIFork, the smart fork which is measuring data about one's eating pace, or SmartyPans, the smart pan which based on information from weight and temperature sensors helps in preparing healthier meals. Besides the mentioned ones, there are also devices such as SmartPlate, TellSpec, SCIO Scanner, Nero-Smart Jar that use mini spectrometers to collect specific information which describes the chemical compounds in the food. Other devices, such as Lumen, use different type of sensors like a CO2 sensor and flow meter to determine the carbon dioxide concentration in a single breath that indicates the type of fuel your body is using to produce energy.

Apparently, this information is not perceived as being very sensitive by its users and is more freely shared, with food photos for example or evaluation of meals in social networks.

Devices that monitor sports activities: There is a wide range of smart devices targeting different kind of sports and body performance assessments. Concentrating on the measurements one can take while doing sports activities, these can be divided in measuring body functions (like heart rate or oximetry that have already been described above), time (time spent, repetitions of an exercise and so on) and location (track, route etc.). In general, the data gathered by these devices is very specific for the sport activities being executed and therefore seem less promising in terms of the contribution that it may provide to the official statistics. One exception would be devices for running and cycling, which are popular activities people choose when trying to improve their personal health. These devices are thus more spread in society and could be interesting for studies on sports, health practising and mobility in general.

First, we will focus on the particular gadgets for cycling as this group is larger than for running. On the market are smart bike systems that, accompanied with the mobile phone applications, provide various functionalities such as navigation, security, lighting, personal assistance etc. Such examples are SmartHalo Bike and COBI.Bike systems. Besides these ones, there are also smart watches, smart pedals and even smart helmets. Although these devices provide different services, they have at least one feature in common – the possibility to inform riders about their fitness performance achieved during the activities, using the data from sensors (built-in and/or belonging to the other devices they are paired with) that measure biking metrics (such as average speed, time, elevation, distance, burned calories).

When it comes to the applications, that cover both scenarios – running and cycling – the most popular one is Strava. Widely used not only for these two, but also for other sports, it can be synchronized with various devices (phone, GPS watch, heart rate monitor). Similar to Strava are Endomondo, Garmin Connect, MapMyRide, TrainingPeaks, Runkeeper, Runtastic, PolarFlow applications, that are also compatible with a wide range of sensors, activity and fitness trackers. All of these applications are available for the IOS and Android platforms and are interconnected in the sense that the data collected, by using one application, can be exported to other applications. The apps that come with the above mentioned gadgets are compatible with some applications such as Strava or Garmin Connect and data can be exported or in other ways utilized, whenever the user personally allows it.

According to their privacy policies, the majority of the data owners offer their users the possibility to share collected data with third parties in different ways. Regarding other users, data can be shared publicly (example: Strava Global Heatmap[16]), due to legislation for service(s) or fee.

From the vast list of gadgets/apps for cycling and running we have found that:

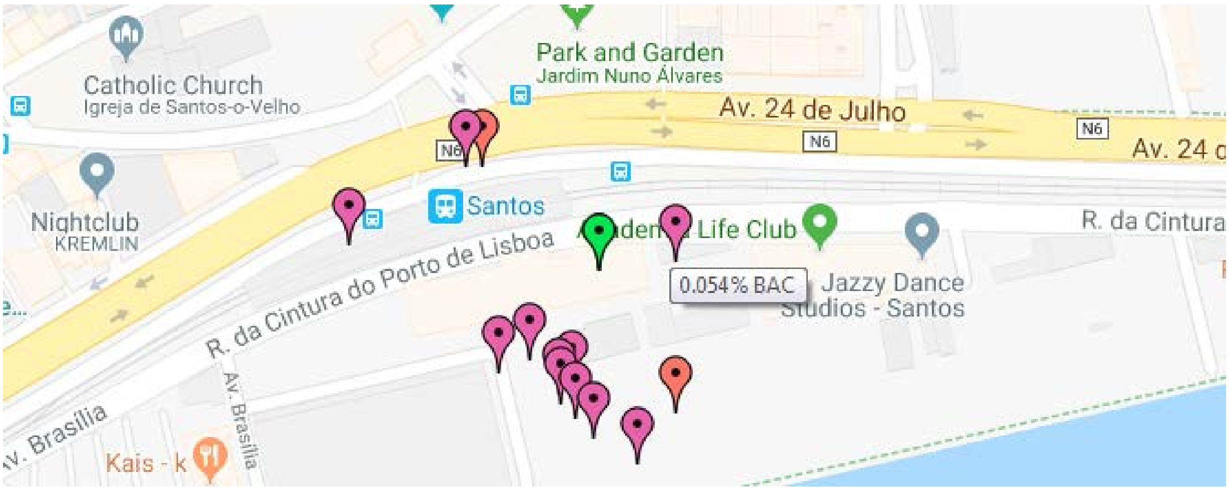
- Strava makes their heat maps publicly available and could make it available for research in an aggregated or anonymized way.
- Endomondo may share data with third parties for research, but their terms of use are not clear.
- Garmin Connect may share or sell activity data to third parties.
- Runkeeper and Runtastic may share their data for research purposes.

Devices that measure blood alcohol level: Probably the most reliable way for measuring the alcohol level in blood is by using the blood tests. The use of blood tests is limited, because they can be carried out only in medical institutions. Other alternative ways for getting similar information that are more convenient, painless and available on a more frequent basis are breathalyzers and portable alcohol trackers. These are devices that are used for measuring and reading the level of blood alcohol content/concentration (BAC). They are the main tool for estimating the level of intoxication and are

[16] There is one sucessful experiment or reverse engineering from the Strava heatmaps back to tabular form data.

regularly used by police officers. Meanwhile breathalyzers, which can connect with the smartphones or other smart devices, provide a different and more personalized experience (Floome, BACtrack), like estimating the recovery time to be within the legal limit for driving, suggesting nearby eating places or even finding you a cab. There was also an improvement on design (*Lapka* breathalyzer) and functionality, which may help the devices to become more widely spread in the near future. Most of the breathalyzers on the market have fuel cell sensors which provide better quality and more accurate results (comparing to results from semiconductor sensors). More discrete and more socially acceptable are wrist wearable devices that are not yet available for commercial purposes (BacTRACK SKYN, Tally, ION^{TM}). While breathalyzers use breath, the wearable devices use skin for an estimation of the blood alcohol level and also give results faster. Both, breathalyzers and wearable alcohol trackers, are compatible with selected smart Android and IOS devices.

Although this information has high sensitivity in some cases like BACtrack, the users who opt-in to share their results publicly can also see other users results through worldview. In the long run, this could be promising towards European Statistics on alcohol consumption.



2. ***To facilitate treatment:***

The smart devices that facilitate treatment have usually three parts: physical sensors with which to register data from the patient and also from her/his surroundings; computational capacity to store, send and receive and sometimes analyze the data and finally the means to deliver either automated actions or actionable advice tailored to the end-user in question. The data analyses can be rather complex, including machine learning and is commonly performed in central servers that then deliver the computational results back to the device. In this group, the medicine dispensation devices receive more attention as the devices for performing smart surgeries, which for example do not produce interesting data for official statistics. The medicine dispensation smart devices are naturally becoming more pervasive in aging groups.

Devices that monitore asthma and inhalers: For treating bronchial asthma. Patients with asthma need two types of inhaled medication: to prevent attacks, they take inhaled corticosteroids that dampen the inflammation that drives the disease. And when symptoms such as coughing or wheezing begin, they use bronchodilators, known as reliever or rescue medication. For now, smart inhalers record only actuation — a time and date stamp of when a patient uses their inhaler. Several companies are developing products that clip onto existing inhalers, like Propeller Health and

Adherium. However, there is the perception that although for now smart inhalers just use Bluetooth technology to detect inhaler use, to remind patients when to take their medication and to gather data to help guide care, they have the potential to improve patients' adherence to asthma therapies and keep their condition under control. So there is already some value for official statistics on the data reporting the usage and it is also an area to keep an eye on for the long term as it relates closely to the air quality and the life conditions.

Devices that measure injections: For the treatment of diabetes mellitus several smart syringe pens are already available on the market, some examples are slim X2 insulin pump with Dexcom G6 Continuous Glucose Monitor (CGM) system, Gocap, InPen and Esysta Bluetooth insulin pen. All these devices have the capacity to register the dose dispensed with a timestamp, but also to continuously register the glucose level in a given patient. These devices are becoming broadly adopted as they are less painful for the patient, significantly reduce the risk of intramuscular injection and provide better data to the physician and the healthcare team. While they provide these benefits for the individual patient, they can also give a contribute to health statistics given the large incidence of diabetes in European population and the interest in this trend evolution as correlated with aging societies.

Devices that measure medication ingestion through smart pills: Although some technological advances (e.g. an electronic medication container lid) can provide information related to medication adherence, they do not provide direct evidence of medication ingestion[17]. Adherence is the degree to which a patient follows medical advice, most commonly with regard to taking medications. Nonadherence to prescribed medications is a significant problem across all medical fields. Among patients with some disorders (e.g. schizophrenia, diabetes, asthma), nonadherence is the largest driver of relapse and hospitalization. This is a very important factor contributing to the spread of smart pills and smart blister packs presented bellow.

However, the information gathered through the use of smart pills raises ethical questions[18] related not only with its subsequent use for statistical purposes, but even medically, for patients who are treated and monitored with these devices. These smart drug-devices hold implications for autonomy and informed consent, device-related "therapeutic misconception", data management, undue influence, privacy and confidentiality. For these reasons, it is not likely that their use for statistics will occurr in the short term outside the boundaries of clinical trials. However, several pills already exist to combine medication with an ingestible sensor (Proteus Digital Health and ID-Cap)[19] or to deliver intestinal injections through robotic pills (Rani Therapeutics)[20].

The most common and freely available automatic pill dispenser or electronic medication containers on the other hand are available and affordable to the public, but most of them are not connected to the internet. Some of the devices are already connected as is the case of GMS WiFi Automatic Pill Dispenser or e-pill MedSmart PLUS and are thus more promising for future statistical use cases.

[17] Inspired by https://www.wearable-technologies.com/2018/05/how-smart-pills-could-revolutionize-healthcare

[18] The Ethics of Smart Pills and Self-Acting Devices: Autonomy, Truth-Telling, and Trust at the Dawn of Digital Medicine, Klugman et al Pages 38-47 | Published online: 20 Sep 2018 https://doi.org/10.1080/15265161.2018.1498933

[19] The pill emits a weak signal when the medication is ingested (the stomach acids act as a battery for a few moments before those same acids dissolve the sensor

[20] Able to navigate through the stomach and enter the small intestine where it delivers an injection without exposing the drug to digestive enzymes

Devices that measure medication ingestion through smart blister packs: Adressing the same market, which is to ensure medication adherence, but using a different type of device, there are also smart blister packs like Med-Ic Smart Blister, which monitor one unit dose at the right time and at the right date.

Besides being available for use with Android and iOS, it is designed for use with blister packaging formats and can include on-board temperature sensing, tracking for sensitive medications, such as monoclonal antibody drugs and even evidence of. Other examples like the smart blister pack that Schreiner MediPharm and Dutch technology company ECCT (Confrérie Clinique) have developed, are already being used with clinical trial participants. However, due to the data sensitvity and the small market expression of these products they do not seem promising for official statistical production just yet.

Concluding remarks on smart health and fitness devices

Devices for healthcare and fitness have very similar characteristics, not only regarding the data itself, but also their sensitiviness and ethic questions that surround their use. The main difference in this case is not related to the nature of the data, but to the way of acquiring the data.

Smart devices for fitness typically store their data with the vendor and reports are sent to the user usually in some aggregated format. Access to the raw or primary data is therefore not easy. There are initiatives like Shimmer that make it easy to pull health data from popular third-party APIs (Application Programming Interfaces) like Runkeeper, iHealth and Fitbit, among others. It converts that data into an Open mHealth compliant format. The Open mHealth was built to structure health data, in order to help companies, organizations and individuals to exchange data and reuse code, making the data easier to understand. The Open mHealth uses schemes to define the structure of health data. The efforts to come up with simple, extensible and clinically valid schemes for the most common and important types of data in healthcare were developed by top clinical experts, data scientists, developers and software architects working together. This kind of harmonization that already exists in this field apparently supports the idea of the data from health devices being not only interesting on itself for the production of official statistics, but also mature enough, at least across some vendors or using third party application like Shimmer, to be explored in more than one platform.

When we look at healthcare smart devices, instead of having the data either on the vendors or in the users possession, data may also be on hospitals or healtcare facilities. This can work either as an advantage or as a disadvantage depending on the cooperation achievable with the healthcare providers. Due to the elevated number of clinical trials currently being carried out across Europe, there is the potential to develop partnerships with doctors/hospitals and health ministries to further explore this data and ascertain its value for specific statistics on diabetes or cardiac disorders for example.

3.4 Smart devices for mobility

Smart devices for mobility are increasingly common. The potential of the data collected by these devices is enormous. The use of the data produced by them enables smart management of both public and private transport. An efficient transport system produces several benefits: it saves fuel consumption, reduces emissions of pollutants, reduces travel time needed to move people around the city, improves the quality of life for citizens with huge economic, environmental and social benefits. The true added value of smart devices for mobility is achieved through their integrated use. Below we will list the main sensors used in public and private transport and their possible use for official statistics.

Smart Traffic Light: The smart traffic light is a special type of traffic light which, using vehicles sensing sensors, can keep the running time on the intersection road higher, with greater traffic flow at that time. The integration of several smart traffic lights combined with public transport detection systems can be used, for instance, to prioritize public transport vehicles by adjusting the traffic lights so that the buses reach the green wave during the entire journey. This would allow people to move faster, reduce fuel wastage and could encourage people to opt for public transport, further reducing road traffic. There are several types of sensors that allow traffic lights to quantify road traffic; the main ones are: (i) magnetic sensors: a loop of electric wire embedded in the asphalt in which a weak current flow allows to detect the presence of metal (vehicle) above it; (ii) optical sensors: typically a camera or infrared sensor that detects the presence of vehicles; (iii) ultrasound sensors: devices capable of detecting vehicles by sending/receiving ultrasound. The sensors are connected to the semaphore control unit, to which they transmit the information collected via cable or via radio. The data received in this way, besides being useful for a traffic signal self-regulation, can be collected to create traffic statistics.

Smart detection cameras[21]***:*** It is a combined system of video cameras and software that allows you to transform normal Internet Protocol (IP) based cameras into smart devices for the detection of a series of data and information on traffic flows and more. These systems allow integrated and customized video analytics solutions for real-time traffic monitoring. The features offered concern not only traffic management but also road safety. These systems are able to: (i) detect stationary vehicles or vehicles in the wrong direction; (ii) detect queues and slow down vehicles; (iii) detect smoke in the tunnel; (iv) make vehicle counts and classifications; (v) track people/vehicles; (vi) estimate the speed of vehicles and the distance between them; (vii) count vehicles; (viii) classify vehicles by size; (ix) detect abandoned objects/lost loads; (x) offer traffic management solutions. The smart detection camera systems are able to handle large amounts of images and data in a reliable manner. They allow road network managers to deal with the criticalities and the continuous evolution of traffic in real time. Moreover, the video analysis algorithms allow to acquire a series of significant statistical data useful during the analysis of traffic flows. A particular type of smart detection cameras are the ANPR (Automatic Number Plate Recognition) cameras. It is a combined system of IP-based cameras and software with the functionality of reading license plates and Optical Character Recognition (OCR) codes for hazardous goods transport. They are special products able to perform reading even at high speeds; they are therefore suitable for both urban and suburban motorways. There are two types of license plate reading technologies: (i) PC basic system or LPR (License Plate Recognition); (ii) ANPR cameras. In the LPR systems, the cameras transfer the images of the vehicles to the personal computer in which the OCR software transforms the text of the photos (the license plate numbers) into characters. In order to use this technology, the cameras must transmit at least 25 images per second, have an infrared illuminator and an anti-dazzle headlight filter. Instead, in ANPR systems, processing takes place inside the camera; OCR software is embedded on the device. This type of cameras are also equipped with an internal database to store the

[21] Inspired by https://www.aitek.it/en/video-analytics-aivu-smart-aid/

detected license plate numbers. Moreover, the ANPR cameras are able to self-regulate in order to ensure the best image quality in any kind of environmental context.

A very interesting example of the use of smart cameras for traffic measurements can be found in the citizen science project *Telraam*[22] in Leuven, Belgium. Citizens attach a low resolution camera to the front window of their homes and open source Python software - created running on a raspberry pi[23] - calculates hourly traffic counts of pedestrians, bicycles, cars and trucks. The smart software analyses and processes the images from the camera immediately so that they are never stored which prevents privacy problems. An opinion was requested by the Belgian Data Protection Authority (DPA). They raised no objections to this method of processing the camera images. The software was created and is made publicly available by the Transport and Mobility Institute (TML) in Leuven. Figure 1 shows the measured traffic intensity based on these data for a part of Leuven.

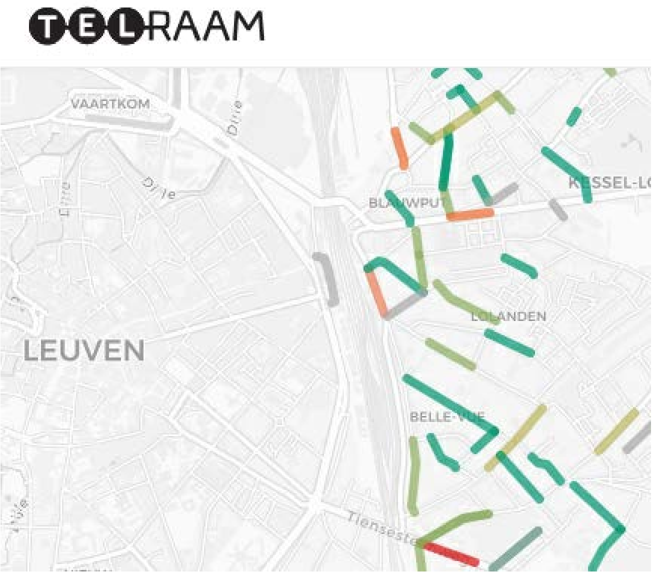


Figure 1: Telraam traffic count project

Smart parking sensors*[24]*: In large cities, finding a parking space is very difficult and therefore creates a waste of time, increased pollution, waste of fuel and stress for drivers. Making the search for a parking space more efficient is one of the most important challenges in order to turn a city into a smart city. The use of parking sensors makes sense only in an integrated system with a software that uses sensor data to help drivers find parking by processing the data and communicating parking lots available via apps. In smart parking systems, the first objective is to ensure accurate and timely vehicle detection. The smart parking sensors include the magnetic sensors and the vehicle detection algorithms that accurately detect the presence or absence of cars in a parking space. The best sensors provide continuous vehicle detection without missing a parking event. They incorporate a magneto-inductive technology with a very sensitive sensor detecting temperature fluctuations. The sensor includes a built-in radio that communicates to a gateway or base station, too. Usually the parking sensor is equipped with a long-life battery (up to 10 years). The most common communication protocol used by the parking sensors is LoRa. It is a long-range wireless

[22] https://telraam.net
[23] One of the most popular brands of single-board computers, see https://en.wikipedia.org/wiki/Raspberry_Pi
[24] Inspired by https://www.pnicorp.com/placepod/

communication protocol that competes against other Low Power Wide Area Network (LPWAN) technologies such as Narrowband IoT (NB IoT) or Long Term Evolution (LTE). The long range connectivity of LoRa can communicate over 100 km. Smart parking systems could provide interesting data on mobility, use of public space and maybe also input for early economic indicators.

Smart passengers counter[25]: Smart counters or Automatic Passenger Counters (APCs) are widely used in our cities: not just around the city streets, or inside public buildings or shopping centers, but on means of public transport, too. The technology used by smart counter is a camera with an embedded software that is able to count people, distinguishing bikes, luggage, adults and children. The system is 98% accurate and provides both real-time and historical data for big data analytics. On buses, trams, trains, subways etc., the cameras are installed over each door and are linked to a counting unit running software, which detects entries and exits. Counts can both be stored locally or transmitted to a server. Users can log into the system securely over the internet to see real-time passenger information. Automatically count passengers getting on and off a bus, train, light rail or other transportation systems provide valuable information to manage the public transport system of a city in a "smart" way. There are many benefits to automatically counting passengers, from cost to operational reasons; some of them are listed in the following: (i) increasing the fare revenue produced by cross-checking passenger numbers with issued tickets; (ii) obtaining information for scheduling, forecasting and service-related decisions; (iii) analyzing key performance indicators (passengers per mile, cost per passenger etc.); (iv) recording times spent waiting at each stop; (v) monitoring passenger numbers over time. Furthermore, an automatic passenger counting system on its own provides valuable information, but when you add vehicle tracking with GPS you can see where a bus, train or tram is as well as how many people are on it. In conclusion, using the IoT to combine all smart devices for mobility, the information available provides insights into people's behavior and helps make cities better places to live.

Smart sensors in vehicles: In the digital age, transport vehicles (cars, trucks, buses etc.) carry dozens of electronic sensors to regulate their functions and monitor performance. The most common sensors used in transport vehicles are cameras, GPS, geo-location systems and laser light to detect and locate objects; most or all of them are smart and connected in the IoT manner. Furthermore, it has been estimated[26] that cars, capable to transmit data via internet, will reach 75% of all cars in 2020 and each of them is able to ship out 25GB of data to the cloud, every hour. That is a lot of information, but what kind of data has been gathered? We can classify the kind of data gathered by vehicles' smart sensors in three groups: (i) data related to vehicle maintenance, with smart cars able to monitor their own components for signs of wear and tear; (ii) data related to driving history: where you have been, the route you took to get there, pit stops along the way, how long you were on the road etc.; (iii) other kind of data such as driving speed, drivers' behavior, road and traffic conditions. This type of data can avoid potential accidents through pre-emptive maintenance. City planners may use information gathered from smart cars in assessing road and traffic conditions, traffic volumes, infrastructure, lighting and safety. In an aggregated form, this data may be used by official statistics to derive indicators on mobility patterns, infrastructure and city smartness.

Car, bike and E-scooter sharing systems: These sharing systems are services, in which bicycles, cars or E-scooters are made available for shared use to individuals on a short term basis for a certain price

[25] Inspired by https://www.retailsensing.com

[26] https://www.gartner.com/en/newsroom/press-releases/2015-01-26-gartner-says-by-2020-a-quarter-billion-connected-vehicles-will-enable-new-in-vehicle-services-and-automated-driving-capabilities

or for free[27]. These systems are widely distributed in many European cities. The two main schemes adopted are: (i) station based and (ii) free floating. In the first one, the people borrow the means of transport from a "dock" and return it at another dock belonging to the same system. In the second, there are no “docks”; people look for a bike, car or E-scooter through a smartphone app, rent it for a certain time span and leave it at the destination. In both cases, the data gathered by the GPS embedded on the means of transport are very valuable. They are used for both public and private aims. For example public transport services, analyzing the traffic, could rethink the frequency of their trains, adapting to the use of the vehicle sharing apps near a station and therefore to the request of the users. Companies could use it for targeted one-to-one campaigns. In the context of shared mobility, these means of transport can be of particular interest as well. Car sharing systems have spread since the early 2000s, thanks to the development and dissemination of the communication systems and the spread of smartphones and their apps that make the system easy to use. Car sharing is also growing in popularity in the context of sustainable mobility and policies to move from possession of vehicles to shared use. Bike sharing systems are increasingly used in larger cities. The E-scooter is not so widespread at this moment but is getting popular especially to facilitate short urban trips. They take up less space than bicycles, are easy to pick up and seem to be growing in popularity for transport to and from bus lines, thus providing a way to bridge the ‘last mile in public transport’. All in all, car sharing, bike-sharing and E-scooter systems use GPS devices that, in combination with data on the use of the systems, could be valuable for official statistics. Statistical offices could use this data for calculating detailed statistics on urban mobility and social networks.

***Portable GPS trackers*:** GPS is used in a growing number of electronic devices, for example in satellite navigators or in smartphones. However, these devices are not sufficiently advanced for activities such as hiking, climbing, swimming etc., because they need the positioning information to be integrated with dedicated maps that show for example hiking paths, mountains and lakes or the sea. The portable GPS overcomes these needs and is therefore extensively used by outdoor enthusiasts such as hikers, fishers or for specific activities such as mushroom hunting. For boating or fishing, there are portable GPS devices that provide integrated nautical charts. Such devices do not only help navigating, they also allow to mark points of interest, for example a dangerous area or, on the contrary, a good place to look for something such as mushrooms. The navigation data can be useful in itself, but possibly even more interesting are the specific marks and annotations added by the people using these kinds of portable GPS trackers. They could be useful for calculating statistics on sports and outdoor activities.

3.5 Smart devices for travel

Technological advances have changed the way we travel. Most travelers plan their trips on the internet, while only a few of them still use travel agents. Moreover, the travel smart devices make travels easier and more organized. We present below a non-exhaustive list, organized in 4 categories:

Smart suitcase: While manufacturers have made advancements in materials and design, suitcases really have not changed much since their modern inception in the early 1900s. But luggage is finally getting smarter, and the options for connected suitcases and related gadgets are beginning to take hold. The ideas are extremely interesting: e.g. being able to check on your smartphone where your bag is, how heavy it is and whether it has been opened. Usually, smart luggage is hard-shelled and

[27] Inspired by https://en.wikipedia.org/wiki/Bicycle-sharing_system

can contain any combination of these features: (i) device charging; (ii) GPS tracking; (iii) remote, app-enabled controls; (iv) bluetooth connectivity; (v) WiFi connectivity; (vi) electronic scales.[28]

Tracking device: Smart tracking devices can help travellers keeping track of their stuff, to find it when something is lost or even to help avoid loosing it***.*** It can be attached to keys, luggage, passport, wallet or anything else important to track its location using a simple smartphone (iOS/Android Compatible) app. If a person cannot find its phone, the smart tracker will get a ping. If the smart tracker is attached to the keys, a person will get notified if it walks out and forgets them.

Smart travel cards: One of the biggest trends in public transportation are smart travel cards. These cards usually contain a chip that can store the card's value. These cards are spread in almost all parts of the world. Contactless smart travel cards are very simple to use. The "contactless" functionality of a smart travel card allows it to make a wireless connection with a reader. The card owner simply holds the card up to a reader and the fare is paid. Contactless smart travel cards, together with the fare payment network behind them, deliver high levels of security. The transport card is, in most cases, anonymous, so no one can glean any information about its card owner.

Global Hotspot[29]: Connectivity can be an issue while abroad, especially when it comes to trusting airport WiFi. These devices help people to connect securely and easily while on travel. The start button on the device activates a local WiFi network for a certain period (usually 24-hour) to which one can connect multiple (usually up to 5) devices.

Considering the 4 categories illustrated above, the only one that could be considered of interest for official statistics is smart travel cards. These cards could be used as an additional source to produce statistics on the phenomenon of commuting.

3.6 Other smart devices

Smart waste management systems: Yet another, maybe peculiar, but also interesting type of smart devices can be found in so called smart waste management systems[30]. Not only can they provide data on the waste management process, in certain countries individual households pay per waste unit and the garbage bin has sensors collecting data for calculating the fee to pay. The data that is generated by these devices might be a valuable input to the monitoring of the introduction of circular economy[31].

IO-Link smart sensors: In the field of industry 4.0 a communication protocol named IO-Link is spreading. According to Wikipedia[32] "IO-Link is a short distance, bi-directional, digital, point-to-point, wired (or wireless), industrial communications standard (IEC 61131-9) used for connecting digital sensors and actuators to either type of industrial fieldbus or a type of industrial Ethernet. Its objective is to provide a technological platform that enables the development and use of economically optimizing industrial processes and operations". This technology is expected to make industrial processes more reliable and transparent as it allows monitoring information to be collected from sensors or actuators directly. It supports different and heterogeneous data types in digital and

[28] https://www.lifewire.com/smart-luggage-4156871
[29] https://www.gearbrain.com/7-best-smart-travel-devices-2576097688.html
[30] https://www.postscapes.com/smart-trash/
[31] https://en.wikipedia.org/wiki/Circular_economy
[32] https://en.wikipedia.org/wiki/IO-Link

analog format and thus is an interoperable communication system able to integrate different fieldbus standards, combining the control system with the physical process. This kind of solutions allow the vertical integration of information for the Industrial Internet of Things (IIoT). In addition to the operational data (typical of standard sensors), the IO-Link connectivity also provides functional data (sensor status, diagnostic information etc.) that are collected through the network and sent to a central server for further processing. Interoperability and vertical integration are the two added values of IO-Links compared to systems currently used in industry. These type of devices allow for efficiency management of functions such as: (i) fault detection for rapid intervention; (ii) monitoring of conditions for predictive maintenance; (iii) identification of the fault component with reduction of man-hours needed to replace or repair. The data generated by IO-Link devices could provide interesting insights into the industry performance and thus could be relevant for statistics on economy.

4 The use of smart devices for official statistics

In this chapter, we narrow the approach from the previous chapter. Starting from the long list of smart devices, we select those that we think are valuable to dive in deeper in future European Trusted Smart Statistics projects.

The first observation we make is that it is not easy to make this selection. Although many examples are promising at first sight, it is also clear that this field is developing fast and the actual use of the devices may change unexpectedly, depending on factors that we cannot foresee today. In a way we can only see the very beginning of a new phenomenon. Therefore, we take a practical approach: instead of scoring the list on some criteria that are probably difficult to quantify, we reviewed the list with the aim to identify a small number of use cases that we think could be valuable to dive in deeper in a follow-up project. While doing this, we keep the following simple yet practical criteria in mind:

- As explained in chapter 2, we *see four foreseeable types of data access:* ***direct data access, pushing computation out***, ***manufacturers' portals*** and ***citizen science***. We would like the selected categories to reflect some ***variety*** in terms of these types of data access.

- As explained in chapter 2, legally it makes a difference whether the data collected by smart devices is classified as ***public sector data***, ***private sector data*** or ***community da***ta. Also, in this respect we opt for some ***variety*** and we try to cover each of them.

- We require the categories selected to show at least some ***maturity*** in terms of ***number of users*** and ***intensity of use***. We feel this is an essential element if the data is to be used for trusted and sufficiently representative official statistics even in a first use case experiment.

- We do not want use cases that are specific to one country. We opt for ideas that are - in principle - ***generally applicable to the ESS***.

Reviewing the long list with these criteria in mind, we come to the following picture:

As mentioned before, ***smartphones*** are the most widely spread smart device around the globe. Their use is mature in all countries. The number of integrated sensors and their capacity is growing continuously and combined with smart software they are an ideal candidate for further experiments.

Some categories of ***smart home devices***, such as data from smart thermostats, are promising. It could be beneficial to start talking to the main providers of these devices to see, whether some computation could be pushed out to gain insights in energy consumption patterns of households. The use of smart devices by citizen science projects for outdoor air quality measurement is promising as well. Not per se for measuring air quality, which is usually not the task of a statistical office, but as an example showing the power of combining real time smart device measurements for deriving fast indicators on statistical phenomena.

When we look at devices for ***healthcare and fitness,*** they have in common that they could provide data that are not easy to measure with traditional data collection. This is certainly a reason to justify the attempt to explore their use for the production of official statistics in more depth, however at this point we feel – also taking ethical and privacy considerations into account – that they are not the best to start with.

When it comes to smart devices for ***mobility***, both in public and private transport, the information they gather are also useful for making a city “smart”. Citizens get many benefits from smart transport devices: the smart traffic lights allow people to move faster, to reduce both traffic and fuel wastage; the smart detection cameras make the roads safer; the smart parking sensors reduce the time needed to find a parking space and thus the pollution. Smart devices for mobility, as a whole, are able to detect traffic flows, to monitor how people move, to count waiting times at bus stops, to measure the kilometers traveled by people, the time needed to move from one point to another in the city and so on. From the perspective of official statistics, these devices could be a rich data source. The data could be used for traffic statistics, urban mobility statistics or generally for mobility statistics; the same holds for data gathered by smart devices embedded on public transport. Especially if used in an integrated manner, they can help produce indicators both in the context of mobility and in the social context.

Turning to smart devices for ***travel***, they are certainly very useful for everyday life, but they do not seem very promising for the production of official statistics. Among those analyzed in the previous chapter, the only device that could be quite helpful for official statistics seems to be the smart travel cards.

From the ***other smart devices***, the waste management systems might be a case to dive in to. They could provide new insights into the circular economy. However, the maturity of their use varies heavily per country, so we would not choose these to start with.

All in all, for a follow-up study that dives deeper into the possibilities of smart devices for official statistics, any of the promising cases mentioned above could be interesting, but we think that the following categories deserve a closer look:

- The concept of using ***citizen science smart devices*** - which has already proven to be useful for air quality and traffic measurements - for nearly real-time indicators on certain phenomena.

- The use of the ***smartphone as a smart device*** exploiting the many sensors and intelligent software to derive new statistical information.

- The use of data from smart devices for ***mobility*** such as - but not excluded to - ***travel cards*** to understand travel patterns and other phenomena of mobility.

In the next chapter, we describe these use cases in more detail.

5 Possible use cases for PoCs

In this chapter, we propose three use cases for a possible follow-up project. As explained in the previous chapter, we feel we are only at the beginning of the growing use of smart devices in society. What seems promising today might change completely within a few years. Therefore, we think freely about some use cases based on smart devices that we imagine could be started in the coming years.

5.1 Citizen science smart devices

Citizen science[33] is scientific research conducted, in whole or in part, by amateur (or nonprofessional) scientists, also known as "public participation in scientific research". The first recorded example of the use of the term is from 1989, describing how 225 volunteers across the US collected rain samples to assist the Audubon Society in an acid-rain awareness raising campaign. These days many citizen science projects exist[34], not only in air quality, weather and climate, but also on history, literature, physics etc.

Citizen science in the context of official statistics could be thought of as "voluntary public participation in official statistics" [35]. One example where this has been the case for years is the birds monitoring projects executed by the Dutch Centre for Field Ornithology[36], where volunteers count birds and other species. The results are an input for environmental official statistics. We do not know of many other citizen science projects in official statistics. The idea of this use case proposal is to setup a smart device based citizen science project, to add some new smart statistical "eyes and ears" to the data collection instruments of official statistics. Such smart devices communicate directly with the statistical offices and are provided/sponsored by the statistical office, possibly in cooperation with other citizen science initiatives.

The *Telraam* project mentioned in the long list (see "smart devices for mobility") serves as a reference framework for thinking. A key concept of this project is the smart processing of observations– in this case low cost camera images - to calculate the variables of interest, which are the number and average speed[37] of pedestrians, bikes, cars and lorries. After processing observations the data is sent to the central database where they are further processed and added to the publicly available data portal. The continuous data streams make it possible to show live figures and typical averages per time unit (e.g. hours, weekday, working day/weekend etc.) and changes of averages over time. The smart devices used consist of a low resolution camera and a raspberry Pi and some open source software provided by the project. The fact that the software is published openly facilitates trust. People can verify that the device does exactly what it is supposed to do and nothing more.

The *Telraam* concept can be viewed as an example of data access via pushing computation out and citizen science. Part of the statistical computation is done directly on the smart device. This paradigm is the starting point for this use case. The statistical community could start or sponsor one or more citizen science projects where low cost smart devices do some statistical observations in a privacy

[33] https://en.wikipedia.org/wiki/Citizen_science
[34] See for example www.zooniverse.org/projects, www.citizenscience.org and www.citizensciencealliance.org/
[35] One could argue here that executing a survey is also a form of CS for offstats. A distinction could be that where a survey sample is designed by the statistical office, in CS for offstats people are free to join.
[36] https://www.sovon.nl/en
[37] Average speed is not yet operational but has been announced by the Telraam project www.telraam.net

safe way. One could imagine working together with the *Telraam* project at first, extending the objects that can be recognized, then scaling up to other smart devices with the ability to count other objects. In all cases, citizen engagement is a key success factor. So ideally the indicators the device produces should have some value to the participant as well. We leave it to a next project to further examine this.

5.2 Passenger mobility by purpose[38] via the smartphone

This section was prepared in collaboration with Giovanna Astori (Istat)

The objective of this use case is to train a machine learning tool fed by a smartphone app to recognize and classify the purposes of the trips made by people, through the Points of Interest (POIs), the trip structure and information about the respondent collected when logging in the app for the first time. The final aim of knowing the purposes of mobility is to build sets of indicators to assess patterns and choices related to mobility behavior, to help decision makers and researchers to optimize transport policies, in particular in the urban context.

Passenger mobility is mainly represented by:

- The place of origin/destination of the trips
- The means of transport
- The purpose of the trips

In addition, distance and duration of each trip must be collected or estimated.

Many case studies have been conducted in this field of the statistical and technological research, where machine learning techniques have been mainly applied to focus on the detection of the mean of transport used, in order to reduce the burden of the mobility surveys.

The purpose of the trip instead, is generally asked to the respondent directly; no mention of relevant studies applying data-driven techniques to detect the purpose of the trip was found hereby.

Therefore, we propose as a use case, the development of a smartphone app to collect data on people's mobility with the specific objective to individualize the purposes of the trips by automatic recognition patterns.

The purpose is assumed to be closely related to the type of place that a person moves to. To simplify the experiment, the following categories could be chosen to describe the type of place as a proxy to classify the purpose:

- Work/professional/commuting
- Education
- Shopping
- Leisure/Private affairs
- Other

[38] "Purpose (trip purpose, destination purpose, activity) - Definition: Travel purpose of a trip is the main activity at the destination of a trip" - Eurostat guidelines on Passenger Mobility Statistics (*page 14*) https://circabc.europa.eu/sd/a/94bf136b-4c6b-42bb-a979-bc64a622cbf8/Passenger%20Mobility%20Guidelines%20July%202016.pdf

- (To converge to official definitions for Passenger mobility statistics, the "Escorting" purpose should be added; in the use case, the trip's purpose could be collected directly from the respondent or not considered at all, to focus only on the machine learning tool)

At the moment of logging-in for the first time, the user (respondent) will only be asked once to indicate, for each proxy-category, the places he visits usually and an essential set of information (with respect to privacy issues at a great extent). The information collected with this short questionnaire is used to create a labeled dataset that will be used to train a machine learning model, with the aim to:

- Classify places that are "new" in a mobility routine, by assigning them to one of the proxy-categories.
- Provide dynamic feedback to extend the initial set of proxy-categories.

The app should collect the following set of data for each trip: (i) the GPS coordinates of the points of origin/destination; (ii) the departure time; (iii) the arrival time; (iv) the time of leaving the place; (v) the POIs visited, which can be collected through the Google Maps API; (vi) the user's familiar places collected with the short questionnaire at the first log-in; (vii) any information on places and/or unusual trip-purpose to improve the accuracy of the predictive model. Moreover, to better individuate and categorize the proxy-classification of the purpose/places, in addition to the physical destination with its attributes, two more (derived) variables should be taken into account: the time spent in the place and the time slot in which the user attends the place.

5.3 Travel patterns based on smart travel cards

Travel cards are used in most European cities and are the most convenient way for frequent public transport users. The cards can usually be topped up with travel passes or money (pay as you go credit). In the case of monthly passes they are personal cards that can only be used by its owner. In both flavours ticketing data from the transportation systems of the metropolitan areas is technically available, if not yet de facto.

The traditional techniques to handle this data, both in the dimension of storage, imputation and processing, have proved inefficiently. The computation and analysis of stages in a trip, routes (with passengers leaving one mode of transportation and entering another in the same trip), commuter movements (going to and back from a particular destination) demand more time and resources than it is feasible without changing the paradigm of data exploration.

Due to the sensitive nature of the data, it may not include any home addresses of ticket or travel pass holders. Nevertheless, it should be possible to determine, approximately, the residence areas of travellers, even if not individually, then of groups of users with a similar profile. This shift in paradigm, working at the group level and not the individual level, will also allow statistical offices to address mobility questions without compromising the citizen's privacy.

Using the travel cards per se may not be enough. When dealing with big data sources it is often the case that we need to use more than one source and also not only similar, but different big data sources, to add value and discover useful and richer information. In this case, if the goals include to understand how people move, to know and predict economical and tourist growth of cities, to identify commuter motion patterns in particular and mobility patterns in general, to know the usage of the city and the load exerted in its main areas and to review and anticipate how certain events can affect recurrent mobility patterns, at least weather forecasts should be used and conceivably also events aggregating portals targeting the studied cities.

The study of travel cards must be able to capture big data (structured or semi-structured through APIs and non-structured through webscraping or similar techniques); and channel it to a big data environment, with appropriate filesystems and databases, fed both in batch and through streaming.

Traditionally, data integration can resort to i) virtual data integration and ii) materialized or data warehousing, through a workflow of data transformations called ETL (Extract, Transform and Load). However, when dealing with large volumes of data arriving at high rates and from heterogeneous data sources, none of them is acceptable. Storing integrated and historical data in a centralized data repository becomes unacceptable in terms of the storage resources required (volume). Moreover, capturing rapid data changes generated by sensors becomes infeasible (velocity). Additionally, differences in data representation among distinct data sources must be captured (variety). Finally, the quality of data available from web data sources is usually poor (veracity).

To address these problems, not only storage and data injection have to be adressed, but also new Data Discovery and Data Analysis methods must be adopted. Making sense of data implies detecting patterns either overrepresented or underrepresented, making use of suitable null models for evaluation and significance evaluation purposes. Suitable methods for data clustering and pattern mining must be considered, with particular focus on streams, multidimensional data and temporal data. Methods for detecting clusters, biclusters and triclusters within the data have to be developed or improved. Triclusters are particularly useful for identifying temporal patterns and are promising in use cases of this nature.

Pattern mining, and in particular temporal pattern mining, is fundamental to discover tendencies, correlations and recurring events and patterns. These, combined with machine and statistical learning models, allow us then to predict and anticipate future events. Clustering and space factorization is promising, but applying these approaches and techniques on data streams raises several challenges, specifically the development of online machine learning models capable of continuously integrating new data into learning models. Although such approaches are reasonable understood when considering linear regression, that is not the case regarding more complex methods such as deep learning models, that must be considered and evaluated.

The main aim of data analysis is to derive models that allow us to learn new facts and rules, discover tendencies and also predict behaviours and events. This is a greater challenge if we consider not only large datasets, but also evolving datasets and data streams. Most statistical learning and machine learning techniques were not developed with those challenges in mind, relying often on batch processing. In the same way the individual imputation that is traditionally applied when facing missing records, for example during the stages of a trip, routes (with passengers leaving one mode of transportation and entering another in the same trip) and commuter movements (going to and back from a particular destination) may not be feasible. This type of problems concerning big data imputation methods will be of interest in many cases beyond the travel cards.

The travel cards use case must focus on questions such as:

- Identifying stages in a trip
- Establishing routes (with passengers leaving one mode of transportation and entering another in the same trip)
- Detecting commuter movements (going to and back from a particular destination)
- Perceiving travel disruptions due to poor weather conditions or other reasons, like constructions and maintenance, traffic jams, sport or cultural events in the city

While addressing the above questions, methods and methodologies should be devised to:

- Detect patterns either overrepresented or underrepresented
- Identify temporal patterns
- Apply these approaches and techniques on data streams
- Derive models in evolving data sets
- Develop and test new methods for imputation in big data sources