Integrating Environmental Data, Citizen Science and Personalized Predictive Modeling to Support Public Health in Cities: The PULSE WebGIS

Enea Parimbelli,[1,3] Daniele Pala,[1] Riccardo Bellazzi,[1] Cecilia Vera-Munoz,[4] Vittorio Casella[2]

[1]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy
[2]Civil Engineering and Architecture Department - DICAr, University of Pavia, Pavia, Italy
[3] University of Ottawa, Ottawa, ON, Canada
[3] Universidad Politecnica de Madrid, Madrid, Spain
enea.parimbelli@gmail.com

Abstract

The percentage of the world's population living in urban areas is projected to increase significantly in the next decades. This makes the urban environment the perfect bench for research aiming to manage and respond to dramatic demographic and epidemiological transitions. In this context the PULSE project has partnered with five global cities to transform public health from a reactive to a predictive system focused on both risk and resilience. PULSE aims at producing an integrated data ecosystem based on continuous large-scale collection of information available within the smart city environment. The integration of environmental data, citizen science and location-specific predictive modeling of disease onset allows for richer analytics that promote informed, data-driven health policy decisions. In this paper we describe the PULSE ecosystem, with a special focus on its WebGIS component and its prototype version based on New York city data.

Introduction: the PULSE project

PULSE (Participatory Urban Living for Sustainable Environments) is an international project funded by the European Commission under the Horizon 2020 framework to undertake research and innovation in cities in Europe, the United States and Asia. The project started in late 2016 and has a planned total duration of three years. PULSE is partnering with municipality leaders of five major cities - Paris, Singapore, Birmingham, Barcelona and New York - to collect information from the public health system, remote and fixed environmental sensors, and citizen-operated mobile devices, to develop a system for the management of public health policies in the urban environment. The project will be the first to build an integrated approach to public health challenges in cities.

The percentage of the world's population living in urban areas is projected to increase from 54% in 2015 to 60% in 2030 and to 66% by 2050 (United Nations, 2014). In absolute terms, more than 1 billion people were added to urban areas between 2000 and 2014. It is important to recognize that cities are not just economic drivers for countries, but are the perfect lab for innovation and research aiming to manage and respond to dramatic demographic and epidemiological transitions (WHO, 2016). For this reasons PULSE has engaged in a collaborative dialogue with five global cities to transform public health from a reactive to a predictive system focused on both risk and resilience (Prasad et al., 2016). In terms of public health risk, the project focuses on the link between air pollution and the respiratory disease of asthma (Toskala and Kennedy, 2015), and between physical inactivity and the metabolic disease of type 2 diabetes (T2D). Within PULSE, health risk is understood to be a combination of environmental and social exposures (e.g. air pollution, poverty) and human behavior (e.g. a sedentary lifestyle). In terms of public health resilience, PULSE focuses on well-being in communities. The goal is to build extensible models and technologies to predict, mitigate and manage public health problems, and promote community health and well-being, in cities (Badland et al., 2017).

This goal is pursued by establishing an integrated data ecosystem (Fig. 1) based on continuous large-scale collection of heterogeneous data available within the smart city environment. The five partner cities will serve as pilot testbed sites providing data, including information generated by a number of citizen scientists (Irwin, 1995) (150 to be enrolled for each city) who will contribute to data collection during the planned 10-months pilot starting in early 2018.

Each participating citizen will contribute to the PULSE data eco-system using the PULSE Urban Participatory App (UPA) and a set of smartphone connected sensors measuring levels of physical activity, air quality and mobility patterns. The availability of such data will allow PULSE to be a pioneer in the development and testing of dynamic assessments of localized geo-based population data. The project will culminate with the establishment of Public Health Observatories (PHOs) in each urban locality, which will mainly rely on the WebGIS component of the PULSE system to integrate, analyze and visualize data to inform health policy decisions and their evaluation. In the following of the article we present the PULSE system architecture with a special focus on the WebGIS and its current development status.

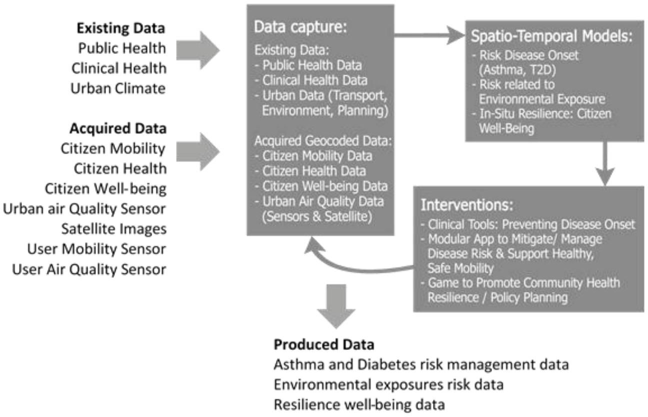
Existing Data
Public Health
Clinical Health
Urban Climate
Acquired Data
Citizen Mobility
Citizen Health
Citizen Well-being
Urban air Quality Sensor
Satellite Images
User Mobility Sensor
User Air Quality Sensor
Data capture:
Existing Data:
- Public Health Data
- Clinical Health Data
- Urban Data (Transport, Environment, Planning)
Acquired Geocoded Data:
- Citizen Mobility Data
- Citizen Health Data
- Citizen Well-being Data
- Urban Air Quality Data (Sensors & Satellite)
Spatio-Temporal Models:
- Risk Disease Onset (Asthma, T2D)
- Risk related to Environmental Exposure
- In-Situ Resilience: Citizen Well-Being
Interventions:
- Clinical Tools: Preventing Disease Onset
- Modular App to Mitigate/ Manage Disease Risk & Support Healthy, Safe Mobility
- Game to Promote Community Health Resilience / Policy Planning
Produced Data
Asthma and Diabetes risk management data
Environmental exposures risk data
Resilience well-being data

Figure 1 - Overview of the PULSE data eco-system. Existing public data, and novel data collected with the help of citizen scientists are combined to build predictive models for disease onset risk and well-being, enabling better data-driven health policies.

The PULSE system architecture and users

One of the key components of the PULSE system is the UPA: the app allows collection of data from citizen participants, some of which are also T2D or asthma patients themselves, through data collection using manual input as well as portable sensors which measure air quality (AirVisual node), activity levels (FitBit Charge 2) and mobility patterns (Smartphone GPS). At the same time users of UPA are not mere data providers. Their use of the UPA comes with the benefit of being able to receive personalized notifications regarding health risk derived from their specific behavior and environmental exposure (e.g. spending time in heavily polluted areas of the city). These notifications are produced by the PULSE analytics/DSS components (see Fig.2) which comprise a knowledge-based PAT DSS and state-of-the-art predictive risk models for asthma (Thomsen et al., 2005; Verlato et al., 2016) and diabetes (Lindström and Tuomilehto, 2003; Rosella et al., 2011). The analytics components take advantage of the integration of big geo-localized community-level data (e.g. hospitalizations due to asthma and prevalence of the disease in specific areas of the city in specific time-frames), the single user characteristics and his patterns of movements around the city. All the geotagged information collected by PULSE is stored in a dedicated repository named GIS DB, specifically designed to effectively manage information which is spatially and temporally oriented. On top of that repository a web-based geographical information system (GIS) has been developed to provide an effective visualization and analysis interface for the PHOs: the PULSE WebGIS. This system is currently under active development and, at time of writing, a prototype working on data for the city of New York is available internally to the PULSE consortium.

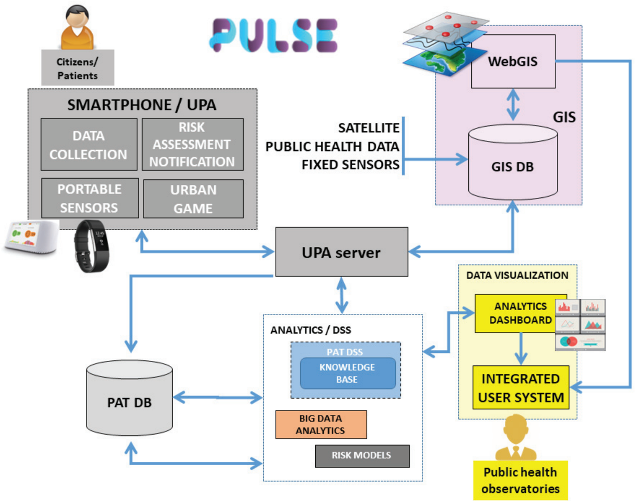

Figure 2 - High-level architecture of the PULSE system and its components. The two main users are citizens using the Urban Participatory App (UPA) and Public Health Observatories (PHOs) using a combination of WebGIS visualization and a more traditional data visualization dashboard.

The PULSE WebGIS component

A GIS is a software designed to store geographic information and tabular data together, in order to jointly visualize, query and analyze them (Chang, 2006). Its aim is to connect the information stored in a database to their spatial reference, allowing the user to explore the geographic features of the data. A WebGIS is basically a GIS that can be accessed through a web browser. Typically, the data contained in a WebGIS can be organized in multiple layers, each representing the information contained in a specific data table. Layers can be switched on and off, overlaid and made partially transparent, thus allowing the user to choose what data to visualize and combine several layers in a new visualization that facilitates data analyses or pattern discovery.

One of the aims of the PULSE project is to collect a large wealth of data from the pilot cities, including maps, orthophotos, satellite optical, multispectral, and hyperspectral images, public health and environmental quality data acquired by air quality sensors, and trajectories followed in space and time by single citizens using the UPA. All of these data have a clear spatial reference and they are stored in the two PULSE data repositories: the patient database (PAT DB) and the GIS database (GIS DB).

The PAT DB stores all the available data concerning the patient personal and demographic information, health status and risk assessment variables, including mobility data from the FitBit sensors, whereas the GIS DB stores data concerning the patients' geolocation and the air quality data, provided by both the fixed air quality sensors and the portable sensors that some citizens involved in the project carry around.

The two databases are connected through the UPA server, that allows data integration with the following effects:

- UPA users are allowed to access health-related information that is usually not promptly available to them (e.g. air quality status in their area);
- The environmental data recorded by the sensors in the areas visited by UPA users, according to the GPS tracking, is added to their personal and health data and used to update their health risk scores, estimated by the predictive models. Personalized notifications are sent to UPA users by the analytics/DSS components.

The PULSE WebGIS is intended to be accessible to all users, some of which have the possibility to download and process the data, according to a role-based access policy.

Every user is allowed to visualize the WebGIS maps of his/her city to consult air quality, demographic and geostatistical maps. These data are represented in different layers that show the data distribution on an interactive map, categorized into the different areas in which every city is conventionally subdivided (e.g. districts or neighborhoods). The air quality data will be continuously updated for the duration of the pilots, while air quality maps will be generated daily and stored in the WebGIS. The users are also able to navigate the information using a timeline, by which they choose specific time frames for the data to be visualized.

PULSE WebGIS prototype: New York City

Currently, a first GIS prototype focused on the city of New York has been developed. The data tables included concern mainly population's general demographics and health-related information, air quality data from the 13 fixed monitoring stations present across town, and the potential correlations between air pollution and health. All the information contained in the tables can be visualized in specific GIS layers, that represent the data distributions according to different city subdivisions. Each layer is referred to the data of a specific table and has color codes and labels designed to highlight a particular phenomenon of interest (e.g. the different asthma prevalence among the city neighborhoods). More specifically, the layers at the current state of the prototype show:

- General census data, such as the population age distributions for each area;
- Asthma/T2D prevalence data for each area referred to the year 2014;
- Asthma/T2D hospitalizations in the year 2014 for each area;
- Air quality monitoring stations' weekly averaged measurements of air pollution parameters, referred to the year 2016;
- Standard satellite and geographic maps from web services such as Google Maps and OpenStreetMap.

The WebGIS prototype has been developed using the open source platform QGIS (QGIS Development Team, 2012). Most of the data have been provided by the New York Academy of Medicine (NYAM), whereas other publicly available information, such as census data, have been found and downloaded from the Internet. Proper ETL procedures are being currently developed to periodically update this data in the GIS DB. Figure 3 presents an example screen of the prototype, in which two layers showing the diabetes hospitalizations in 2014 and the NYC map are activated and overlapped. More data concerning prevalence, hospitalizations and calls to 911 referred to more recent times is being acquired, and more air quality data will be provided by the portable sensors as the pilot studies proceed.

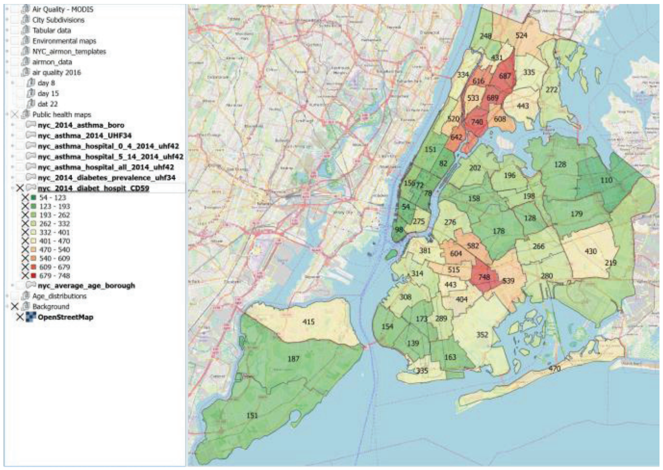

Figure 3 – Example image from the PULSE GIS prototype. 2 layers are active and overlapped: one showing a geographic map of New York City (OpenStreetMap) and the other showing the count of diabetes-related hospitalization in the year 2014, in each of the 59 community districts of the city. Each polygon has a label, and associated color gradient, that shows the number of hospitalizations. The left menu dynamically controls layer visibility.

The possibility to download data tables and visualize the layers allows to analyze the data both visually and analytically to discover new, potentially relevant, risk factors for the predictive models to use and make new geostatistical considerations. Some examples which are being considered in PULSE are:

- Find the correlation between air pollution and asthma looking at the number of asthma hospitalizations increase in the days with poor air quality;
- Investigate the possible correlations between the population age and/or race distribution and the differences in asthma and diabetes prevalence in the diverse areas (Center for Disease Control, 2017);
- Find possible correlations between the population economic and cultural status and the prevalence of the 2 pathologies (i.e., does the economic status of the poorest/richest areas influence the life habits of the population, thus changing their exposure to risk factors?);
- Find possible correlations between the population cultural and economic status and the number of hospitalizations in each area (i.e., does a person's cultural and economic background influence his/her tendency to call 911 when he/she is not well?).

Newly identified significant correlations can be jointly analyzed with the citizen personal data and other risk assessment variables available in literature, to enrich and re-calibrate the predictive models used by PULSE making them more specific for the local context of NYC. These variables can be also included in the PULSE PAT DSS, in order to generate personalized notifications to be delivered to the UPA citizen users, at the same time allowing the PHOs to design, deliver and evaluate more accurate and effective interventions.

Conclusions

In this paper we have presented the PULSE system with a special focus on its WebGIS component, highlighting how the collection and integration of georeferenced data in a more comprehensive data ecosystem could benefit both citizen users of the system (which act both as data-provider and system-generated recommendations consumers) and public health services. Preliminary exploratory data analyses enabled by the prototype developed for NYC show how the discovery of novel correlations between location-specific risk factors and relevant health outcomes (such as disease prevalence and hospitalizations) is facilitated by the WebGIS. Further results from the other PULSE pilot studies are needed to confirm the discovered preliminary hypotheses and to develop context-specific, personalized risk models which consider both global risk factors and location- and time-dependent environmental exposure as well as behavioral factors to improve T2D and asthma onset prediction in the urban-dwelling population.

Acknowledgments

The authors want to thank J. Pagan, E. Fisher and K. Madondo from the NYAM for providing the data used by the WebGIS prototype for the city of New York. This work has been funded by European Commission under grant agreement GA-727816.

References

Badland, H., Foster, S., Bentley, R., Higgs, C., Roberts, R., Pettit, C., Giles-Corti, B., 2017. Examining associations between area-level spatial measures of housing with selected health and wellbeing behaviours and outcomes in an urban context. Health Place 43, 17–24.

Center for Disease Control, 2017. Asthma - Most Recent Asthma Data. https://www.cdc.gov/asthma/most_recent_data.htm (accessed 11.9.17).

Chang, K.-T., 2006. Geographic information system. Wiley Online Library.

Irwin, A., 1995. Citizen science: A study of people, expertise and sustainable development. Psychology Press.

Lindström, J., Tuomilehto, J., 2003. The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes Care 26, 725–731.

Prasad, A., Gray, C.B., Ross, A., Kano, M., 2016. Metrics in Urban Health: Current Developments and Future Prospects. Annu. Rev. Public Health 37, 113–133.

QGIS Development Team, 2012. QGIS Geographic Information System. Open Source Geospatial Foundation Project.

Rosella, L.C., Manuel, D.G., Burchill, C., Stukel, T.A., PHIAT-DM team, 2011. A population-based risk algorithm for the development of diabetes: development and validation of the Diabetes Population Risk Tool (DPoRT). J. Epidemiol. Community Health 65, 613–620.

Thomsen, S.F., Ulrik, C.S., Kyvik, K.O., Larsen, K., Skadhauge, L.R., Steffensen, I., Backer, V., 2005. The incidence of asthma in young adults. Chest 127, 1928–1934.

Toskala, E., Kennedy, D.W., 2015. Asthma risk factors. Int. Forum Allergy Rhinol. 5 Suppl 1, S11-16.

United Nations, 2014. World Urbanization Prospects: The 2014 Revision, Highlights. Department of Economic and Social Affairs. Popul. Div. U. N.

Verlato, G., Nguyen, G., Marchetti, P., Accordini, S., Marcon, A., Marconcini, R., Bono, R., Fois, A., Pirina, P., de Marco, R., 2016. Smoking and New-Onset Asthma in a Prospective Study on Italian Adults. Int. Arch. Allergy Immunol. 170, 149–157.

WHO, 2016. Global Report on Urban Health. http://www.who.int/kobe_centre/measuring/urban-global-report/en/ (accessed 11.6.17).