Exploring the relationship between mobile phone data and wastewater flows: evidence from five Swiss catchments

Jörg Rieckermann[1]*, Andy Disch[1], Nicolas Neuenhofer[1], Stephan Baumgartner[1,2], Christoph Ort[1]

[1]Department of Urban Water Management, Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland
[2]Institute of Civil, Environmental and Geomatic Engineering, ETH Zürich, 8093, Zurich, Switzerland

Correspondence to: Jörg Rieckermann (joerg.rieckermann@eawag.ch)

Abstract

Emissions from urban drainage systems can have unwanted consequences for human and environmental health. Unfortunately, traditional water quality monitoring in sewers is expensive and not comprehensive enough to provide detailed data on pollution across an entire catchment. However, with the increasing digitization of society, alternative data sources such as mobile phone data offer new opportunities to assess wastewater production and dynamics. In this study, we investigate the relation between mobile phone data and wastewater flows in five catchments in Switzerland with different characteristics and sizes, using data from the largest Swiss telecom provider and simple multiple linear regression models. The initial results of this study are promising, although the degree of correlation observed between mobile phone data and wastewater production is rather low (max. R^2=0.73) and varies greatly from catchment to catchment. As expected, we find non-linear effects in the data and more advanced modeling approaches, e.g. considering flow distances or dynamic wastewater travel times in the sewer network, may be needed to develop reliable predictions. In addition, we find that privacy protection issues currently limit the applicability in small catchments. Thus, we expect that the mobile data will benefit from domain-specific preprocessing for wastewater applications. The potential applications of this approach are far-reaching, including applications in urban drainage, wastewater treatment, drinking water, wastewater-based epidemiology and climate change adaptation. Last, but not least, wastewater flow or pollution data could even improve the data quality of the mobility data.

Keywords: Mobile phone traces, Wastewater generation, Flow, Sewers, Digitalization, Monitoring

Highlights:

- This study presents the first high-resolution dataset of mobile phone-based population indicators with hourly data and corresponding measurements of sewer flows.
- It marks the first application in sewer hydraulics, with a comparison across several catchments of different sizes.
- The research highlights lessons learned regarding the limitations of current mobile phone data and simple linear models which predict flows from mobile phone-based population indicators and discusses ways forward.
- Based on our findings, we discuss broader implications of mobile phone -based population estimates for urban water management.

Introduction

Sewers are an essential component of modern urban infrastructure, designed to transport wastewater and sewage away from homes and businesses. However, the discharge of untreated or poorly treated wastewater into the environment can have serious consequences for human health (Brokamp et al., 2017; Li et al., 2023; McGinnis et al., 2022) and the environment (Pistocchi and Dorati, 2018). Sewers are a major source of environmental pollution, and this issue is exacerbated by several factors.

One of the main challenges in quantifying the pollution loads in wastewater systems is the difficulty of accurate water quality monitoring (Lepot et al., 2016). Traditional sensors are often expensive and difficult to install, and grab sampling is often not adequate to capture the pollutant variability during dry and especially wet weather (Gruber et al., 2005, 2019).

To monitor pollutant load in wastewater, it has been suggested i) to predict wastewater pollution from related data sources, such as in-line electrical conductivity and turbidity measurements (van Daal-Rombouts et al., 2013) (see Section 2) and ii) to use non-contact multispectral measurements (Agustsson et al., 2014; Bourgeois et al., 2001), which can be measured with high temporal resolution. However, in-line measurements are often unreliable because of sensor failure and non-contact measurements are not yet available.

Interestingly, another relevant data source could be mobile phone traces, which can track population size and mobility in certain areas in near real-time. This provides insights into the spatial and temporal patterns of urban activity (Atinkpahoun et al., 2018). Such data offer several advantages over traditional flow or pollutant sensors. They are available for an entire wastewater catchment in near real time, and a “soft sensor” developed from this data source would require much less maintenance, power supply, calibration efforts, etc. than traditional sensors. In recent years, mobile phone traces have been increasingly used in research projects, and the availability of this data has opened up new possibilities for understanding complex urban systems (Palmer et al., 2013). One of the areas where the potential of mobile phone data has yet to be studied in detail is the monitoring and prediction of wastewater generation (Thomas et al., 2017).

It seems straightforward to use mobile phone traces to estimate the contribution of each subcatchment to the overall wastewater load, as an increase in population size should, on average, cause an increase in generated wastewater. However, using mobile phone data for this purpose poses substantial challenges. First, the coverage of the service providers causes unknown bias, which might even change during a day due to the mobility of the population. Second, the daily or hourly resolution of the mobility data may not align with the finer time granularity of wastewater generation and production, which happens on the minute scale (e.g. toilet use and household appliances). Third, there is an additional source of uncertainty stemming from non-stationary transport processes. Open channel flow in gravity sewers attenuates peak flows and other disturbances, e.g. from pumping stations (Ort, 2014) increase the complexity of the relationship between mobile phone data and observed flows, especially in large catchments. Fourth, mobile phone traces are often heavily processed due to privacy and data protection issues, which might cause systematic underestimation, especially in small or micro-catchments. Thus, the utility of mobile phone traces in assessing wastewater generation and predicting flows in wastewater systems remains uncertain.

In Switzerland, the use of mobile phone data for wastewater prediction and monitoring has been explored in a very early pilot study (Troxler, 2018). The study focused on a ski race event, where the population increased significantly over several days, providing an opportunity to test the use of mobile phone data to predict changes in wastewater production. The results using data from the provider Swisscom, which has the widest coverage, showed a correlation coefficient of 0.95, indicating a strong relationship between population activity and wastewater production (see SI Section A). However, without a detailed understanding of the pre-processing methods used in the study, it was difficult to assess the accuracy and reliability of the results.

The aim of this study is therefore to predict wastewater flows from mobile phone-based population counts, as a first step towards predicting wastewater generation or pollution dynamics. To achieve this, we used data from the Swisscom Mobility Insights (SMI) platform[1], which provides self-service data for municipalities and has the potential to be used for a range of urban and environmental applications.

This study is novel in several ways. Firstly, it is the first time that mobile phone data has been used to model wastewater production across different spatio-temporal scales in Switzerland. Additionally, it is the first time that the SMI platform has been used for this purpose, providing near-real-time data that can be directly used in the analysis. Furthermore, the data used in this study will be openly available, including information on sewer flows in five (sub)catchments across different spatio-temporal scales, which can be used to provide a more comprehensive understanding of wastewater production dynamics.

The initial results of this study are promising, with correlations observed between mobile phone data and wastewater production for small catchments. For the larger catchment, however, it is important to note that the population density-to-flow relationship is clearly non-linear, and that the different temporal resolutions of the data present challenges in understanding the underlying processes in detail and non-linear modeling approaches may be needed to improve flow predictions. Despite the current limitations, the potential applications of using mobile phone traces in urban water management are far-reaching. Based on the lessons learned in this study, we therefore provide an extended discussion on several urban water management problems from assessing industrial wastewater production to wastewater treatment plant (WWTP) inflow generation, real-time control of drinking water systems and wastewater-based epidemiology.

The remainder of the paper is structured as follows: First, we describe the material and outline the methodology; second, we present the results; third, we discuss the findings; and finally, we conclude the paper with a summary and recommendations.

[1] https://mobilityinsights.scapp.swisscom.com/

Material

The catchment of AV Altenrhein and selected subcatchments

The Altenrhein wastewater association (here: AVA) is a joint association of 17 municipalities from two cantons responsible for sewage disposal, wastewater treatment, and sludge treatment including disposal. The association includes eight municipalities from the canton of St. Gallen and nine municipalities from the canton of Appenzell Ausserrhoden. The catchment area of AVA (see Figure 1) covers 1702 hectares and is largely built-up. The northern St. Gallen municipalities are characterized by a predominantly flat topography, in contrast to the southern Appenzell municipalities with pre-alpine terrain. Drainage is carried out in a 56% combined system and 44% separate system, the latter being mainly found in the Appenzell municipalities. In 2019, 62’915 inhabitants were connected to the Altenrhein WWTP, with additional 20’417 population equivalents (of which 70% residential and 30% industrial) (AVA, 2020).

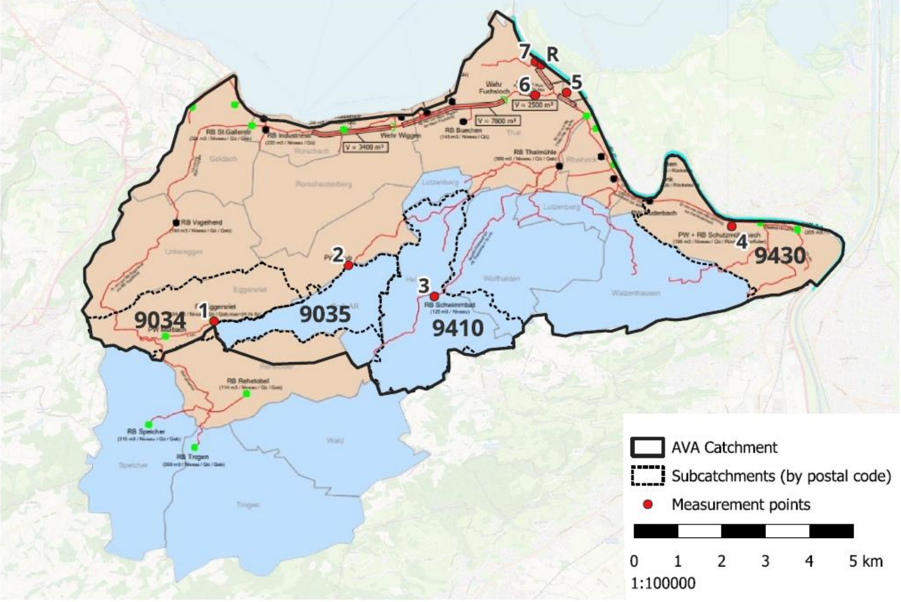

Figure 1: Depiction of the catchment area of AV Altenrhein. R denotes the rain gauge, while 1-7 represent specific flow monitoring sites aligned with catchments closely resembling postal code zones.

Flow and rain monitoring data

As reference data, we use high-resolution flow data in 1-minute resolution from seven monitoring sites in the catchment of AV Altenrhein in the period from July 15, 2019 to September 15, 2019. The sites were selected in close collaboration with the AVA such that the catchment associated with each measurement site corresponds to a geographic area with a single postal code, or a group of postal codes. This was important to extract corresponding mobile phone data (see below).

To assess the relationship between mobile phone-based population densities and observed wastewater flows, we not only included small and large sites, but also selected a period that includes a potentially large population shift: the start of a new school term. Around the start of school, the population increases and generates more wastewater, which can be used to better estimate the effect of mobile phone users on the dependent variable. Therefore, the monitoring period was chosen from July 15, 2019 to September 15, 2019, which comprises 27 days of holidays and 35 days of working days with presumably normal population activity in the catchment (school start: August 11, 2019). In addition, we explicitly chose a time before the COVID-19 pandemic, which started in Switzerland around February 2020. The technical specifications of the flow measurements are summarized in Table 1.

In our study, we accounted for factors beyond population density and industrial activities that influence wastewater flow in sewers. Specifically, we incorporated rainfall data to consider rainfall-runoff and rain-induced infiltration and inflow. To collect accurate and reliable rainfall data, we used a tipping bucket gauge installed at the WWTP. The utility staff considered this gauge to be the most reliable instrument for capturing local, high-resolution rain data. We ensured its precision and reliability by regularly maintaining the gauge and its tipping mechanisms. Furthermore, the collected data underwent periodic quality control processes to maintain data integrity.

The observed flows (Figure 2, top) include mostly dry weather, with 13 rain events or wet periods. The cumulated inflow to the WWTP Altenrhein is the sum of ID 5, 6 and, to a very small degree ID 7. The other monitoring sites describe subcatchments which are located upstream of 5 or 6.

We also observe a clear effect of the working period (to the right of the red dashed line). Before, flow frequencies are lower and peaks more dampened. After the end of holidays, we see increased flow patterns with a generally higher variability. While some signals show the typical shape with a flow peak in the morning and one in the evening, others show more random and frequent variations, which show the activity of pumping stations. Thus, simply through visual analysis of the flow data, we can identify a regime change. The question is whether such changes are also reflected in the mobility data.

Table 1: Data sources provided by AV Altenrhein for high-resolution rainfall and flow data used in this study. R indicates the rain gauge and the numbers correspond to flow monitoring sites where wastewater discharge is assessed using various Venturi installations and bubbler level sensors.

ID	Name	Type of sensor	Postal Code	Static population [E]	Area [km2]
R	Regenmesser ARA	Tipping bucket rain gauge		-	-
1	Zulauf Regenbecken Eggersriet	Metal flume, retrofitted	9034	2275	8.9
2	Zulauf Pumpwerk Grub AR	Concrete flume	9035	1029	4.2
3	Messstelle Schlachthof Heiden	Metal flume, retrofitted	9410	4175	7.5
4	Messstelle Töbelibach St. Margrethen	Metal flume, retrofitted	9430	5891	6.9
5	Messstelle Ost	Metal flume, fix installation		62915	108.2
6	Messstelle West	Metal flume, fix installation		62915	108.2
7	Messstelle Altenrhein	Metal flume, fix installation		62915	108.2

Swisscom Mobile Phone Data

Swisscom, the leading mobile service provider in Switzerland, offers anonymized and aggregated mobility data from smartphones through its Mobility Insights service. This data provides valuable insights for various industries, including tourism, advertising, retail, environment, and planning. With six million anonymized SIM cards, for a population of 8.575 million (2019), Swisscom claims to be a reliable source of quality movement data.

A positioning algorithm is utilized to determine the location of SIM cards, which are aggregated into 100 x 100 meter tiles. To ensure compliance with data protection regulations, Swisscom only reports data if 20 SIM cards or more are in an area. Technically, the process of automatically extracting so-called "network events" from mobile communication systems occurs overnight and involves replacing IMSI numbers with hashes to protect privacy[2].

Specifically, we use such population densities from 17 postal codes, which were composed of the tiles corresponding to the sub-catchment area upstream of the flow metering location. Technically, the data were processed by Swisscom using shape files of the relevant postal code areas.

[2] https://www.swisscom.ch/en/business/enterprise/offer/platforms-applications/data-driven-business/mobility-insights-data.html

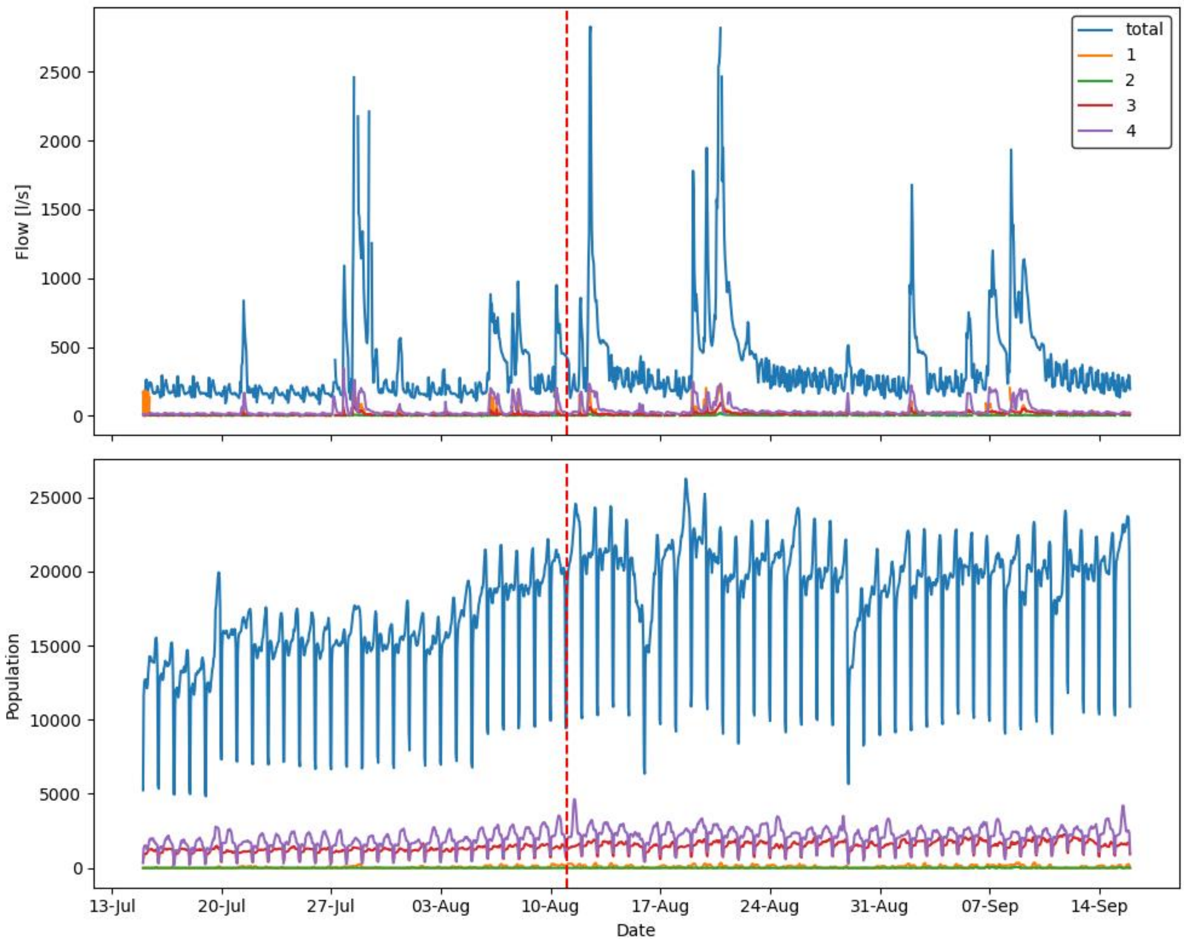

Figure 2: (Top) Illustration of flow monitoring data in the four subcatchments and the total catchment of the AVA. (Bottom) Network events/people connected in a postal code over time. A higher number of people can be detected for the period after the summer holidays (to the right of the red dashed line). The population count is estimated by automatically extracting network events from the Swisscom mobile communication systems. If there are less than 20 SIM cards in a grid cell (100m x 100m), SMI reports 0 counts.

Figure 2 (bottom) presents the population size in each area over time for the 17 postal codes. We see that, similar to the flow data, some postal codes contribute much more than others do and for some we do not get any counts. This is due to the privacy limit, which ensures anonymity in areas with lower population densities.

Methods

Pre-processing of mobile phone and flow data

The mobile data and flow data sets underwent several preprocessing steps. First, the flow data were resampled to match the temporal resolution of the mobile data. Second, the days disturbed by rainfall-runoff were identified through visual inspection of both the rain gauge data (R) and flow data (1-7) and subsequently removed from the dataset. Third, data collected during weekends are excluded. Next, upon visually examining the population data, irregularities were noticed during the midnight hours. As a result, the data recorded between 11p.m. and midnight were eliminated. This process yielded a total of 22 days with dry weather observations. Finally, both population and flow data sets were aggregated to lower temporal resolutions of three, six and twelve hours to investigate the impact of different temporal resolutions on the data analysis.

Exploratory analysis and regression analysis

The exploratory data analysis performed in this study involves plotting and visual analysis of the data, and comparing a set of performance metrics. To quantitatively describe the relation between flow and population, we performed a regression analysis.

The primary objective of the regression model is to establish a predictive relationship between the dependent variable – flow – y_t and the explanatory variable – population size – x_t. We choose a simple time series regression framework, where we incorporate lagged values of the explanatory variable due to temporal dependencies originating from the travel time in the sewer. The β_1 coefficient reflects the instantaneous, depending on the temporal aggregation, influence of the current population value. The coefficients β_2 through β_5 capture the effect of the respective lagged population counts. The model acknowledges the existence of unaccounted variability and stochastic factors by introducing the error term ε.

$$y_t \sim \beta_1 * x_t + \beta_2 * x_{t-1} + \beta_3 * x_{t-2} + \beta_4 * x_{t-3} + \beta_5 * x_{t-4} + \varepsilon$$

This model formulation follows the traditional time series analysis approach, seeking to reveal underlying connections between variables over consecutive time periods. It facilitates an exploration of how current and past population counts (x_t through x_{t-4}) influence the progression of the flow (y_t), thus enhancing a thorough comprehension of temporal dynamics within the examined system.

The determination of the appropriate number of lagged values for inclusion in the model was guided by a meticulous analysis of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots.

These analyses were conducted across various catchments and time aggregation levels to compensate for travel time in the pipe network. After careful examination, it was observed that a lag number of four emerges as the optimal choice, rendering the resultant model particularly robust.

To assess the efficacy of the model in predicting the observed outcomes, the multiple linear regression model was calibrated for each (sub)catchment and temporal aggregation level on an 80% training set. The evaluation of each model relied on the predictive performance on a 20% test set and quantified using the performance metric coefficient of determination (R^2). This was computed with an ordinary least-squares regression:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

Where SS_{res} describes the residual sum of squares

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

Where y_i are the observed flow measurements and f_i the predicted flows based on population density, e_i are the residual errors. SS_{tot} describes the total sum of squares, proportional to the variance of the flow

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$

The R^2 metric was calculated for all catchments and temporal aggregation levels. Note that, at this early stage of this screening analysis, we did not perform a comprehensive statistical analysis, including the holidays, weekends, catchment size, etc. as explanatory variables.

This approach forms the basis of the data analysis performed in this study and enables the investigation of the relationship between different variables, such as rainfall and catchment response, over different time scales. To clean the data, e.g. to filter out disturbing factors such as rainfall, it is important to have high-resolution flow data with a temporal resolution of minutes. This approach allows for a more nuanced analysis of the data, as it enables the examination of the impact of these effects over longer time periods, while minimizing the potential for sampling biases that may occur at shorter temporal scales.

Implementation

We used Python to analyze the flow and SMI data, mainly using common libraries such as pandas, scikit-learn, and matplotlib. We have made the code publicly available at the public repository of the project (see Section **Error! Reference source not found.**).

Results

We present the results for the exploratory data analysis and the numerical experiments for the total catchment of the WWTP (Altenrhein) and the four subcatchments (Eggersriet, Grub, Heiden, and St. Margrethen). The scatter plots in

Figure 3 portray the relation between the population size and the flow within the total AVA catchment and the four investigated subcatchments, using data aggregated to one, three, six and twelve hours.

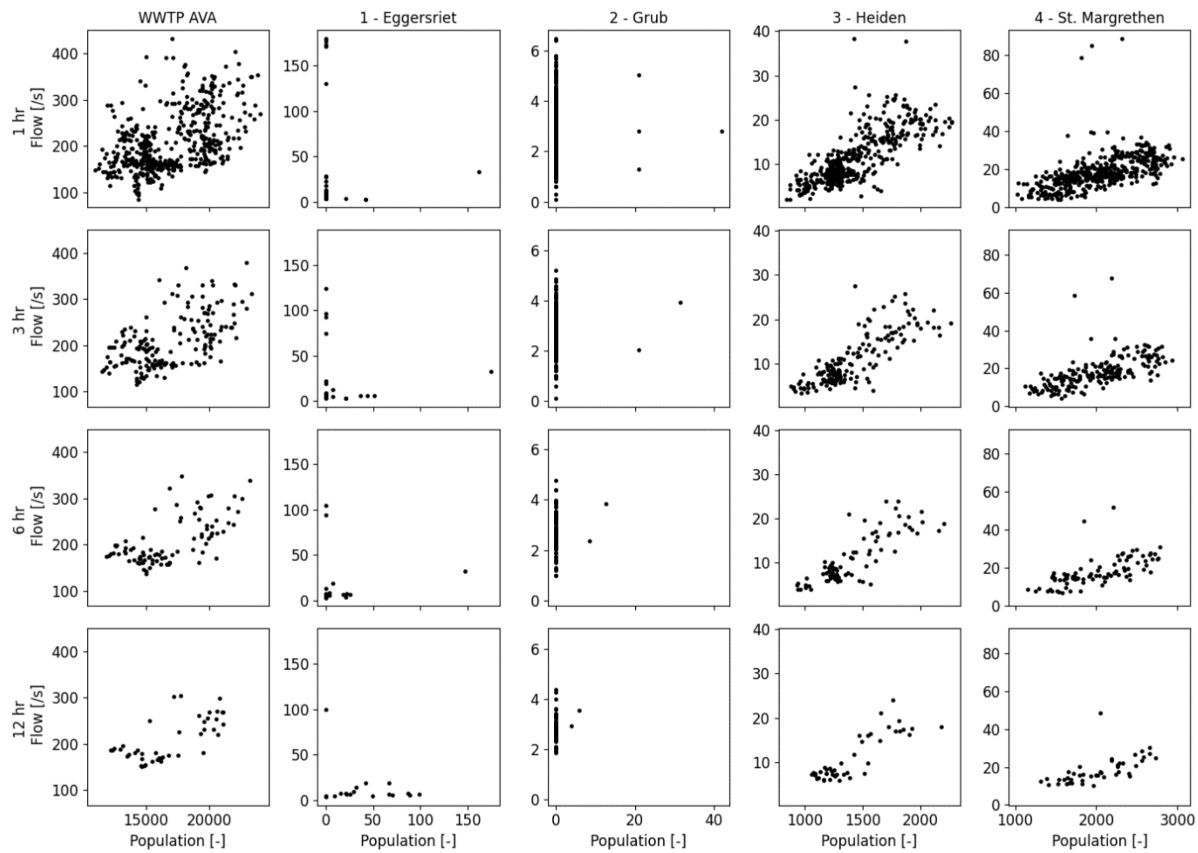

Figure 3: Scatter plots of population vs. flow in the AVA catchment and the four investigated subcatchments for different temporal aggregations.

With the exception of the subcatchments of Eggersriet and Grub, the plots exhibits an overall trend indicating a positive relationship between flow rates and population sizes. Note that this representation only displays the current (aggregated) population versus the current flow and does not include the lagged population counts.

While positive correlations between current population sizes and wastewater flows are observable in

Figure 3, the R^2 of the test sets on the multiple linear regression models vary greatly depending on the (sub)catchment and temporal aggregation, see Table 2. It is initially not intuitive that there should not be high R^2 for a substantial catchment like AVA. One plausible explanation for this phenomenon could be attributed to the varied aggregation levels of the data. While flow is naturally "integrated" over the catchment by the travel time wastewater transport process, the SMI data is not aggregated in time and remains at a more granular or "raw" level. Interestingly, as the temporal aggregation increases to three hours, we see an initial improvement in the R^2, before it drops off again for higher temporal aggregations. However, it is crucial to note that this relationship is not linear, and caution is warranted in the execution of such adjustments, as elaborated in the subsequent discussion.

The effect of pre-processing, particularly the implementation of a privacy cutoff in population data, where population size is reported only above a specific threshold, can introduce noise or even systematic bias into the analysis. This is especially evident in Trogen (a subcatchment not included in this study), where the limitation concerning the minimum population density is visibly apparent at the beginning of the new school semester (Figure SI 2). This approach can lead to a significant underestimation of the actual population, as it results in the exclusion of areas or regions with population sizes below the defined threshold. The severity of this underestimation depends on the catchment population densities and is challenging to predict a priori.

Despite utilizing a straightforward methodology and limited dataset, our discoveries broadly correspond with the anticipated correlation between population density and discharge, even when examined on a 1-hour aggregation level.

Table 2: Results of OLS regression analysis. For each (sub)catchment and temporal aggregation, the performance metric (R^2) of the test set is provided along with the five model coefficients corresponding to the population value (β_1) and the four lagged population values (β_2-β_5)

ID	(Sub)catchment	temporal aggregation [hr]	R^2	Model coefficients				
				β_1	β_2	β_3	β_4	β_5
Total	**WWTP AVA**	**1**	0.24	0.005	0.008	0.005	0.007	-0.015
		3	0.13	0.002	0.004	0.002	-0.005	0.012
		6	0.70	-0.002	0.005	0.003	0.006	0.004
		12	0.73	0.009	0.015	-0.010	0.000	0.006
1	**Eggersriet**	**-**	-	-	-	-	-	-
2	**Grub**	**-**	-	-	-	-	-	-
3	**Heiden**	**1**	0.47	0.015	-0.005	0.006	-0.001	0.001
		3	0.72	0.010	0.003	-0.002	0.001	0.004
		6	0.72	0.011	0.004	0.005	-0.005	0.003
		12	0.28	0.005	0.011	0.004	0.025	-0.024
4	**St. Margrethen**	**1**	0.38	0.006	0.008	0.000	-0.012	0.009
		3	0.30	0.001	0.008	0.004	-0.008	0.007
		6	0.08	0.004	-0.002	0.005	0.001	0.007
		12	-1.18	0.002	0.013	0.015	-0.013	-0.005

Discussion and Outlook

In this study, we found relevant relations between mobile phone data and wastewater flow. However, to fully understand the potential of this approach for addressing real-world problems, the results of this study warrant a broad discussion with regard to topics such as i) data processing, ii) alternative explanatory variables and features, iii) chosen data analysis methods, iv) sewer applications, v) WWTP applications and industrial water use, and vi) drinking water applications. We will also discuss the implications for wastewater-based epidemiology, and improving population size estimates from wastewater flow data. In this section, we will briefly summarize the main points and present a detailed description in the supporting information.

Characteristics and processing of the SMI mobility phone data

The SMI mobility phone data from Swisscom played a crucial role in our study, particularly through the "*Heatmap*" product. Other products like "*Dwell Times*" or "*Origin Destination*" were deemed less relevant for our specific objectives. An interesting observation was a consistent 1-hour dip in the SMI data at the end of the day, resembling the nighttime decrease in wastewater production (Figure 2, bottom). Despite phones remaining active during the night without logging out, technical issues likely caused this dip. To enhance our analysis, we therefore opted to exclude night hours during data cleaning.

The temporal resolution of both data sources presents an open issue, as their simple correlation analysis does not capture the intricate mapping from population size and hidden wastewater generation to wastewater flows. Privacy protection, with a cutoff of 20 SIM cards per tile, poses a significant limitation, especially for less urbanized or small subcatchments. Custom processing by the telecom operator to provide specific data for urban water applications could address this, especially for small catchments. Additionally, Switzerland's mountainous topography may impact mobility data accuracy, with mobile devices connecting to distant antennas, leading to potential discrepancies in location accuracy, especially in mountainous areas.

Considering further alternative data sources to predict wastewater flows

In exploring alternative data sources for predicting wastewater flows, the study focused on correlating mobile phone data with wastewater dynamics. However, simple linear relationships proved insufficient, especially for large catchments. To enhance accuracy and understand the link between population size and wastewater dynamics, incorporating additional data sources is recommended. For instance, traffic data could help mitigate bias in population size, as those in traffic typically do not contribute to wastewater. Challenges arise with drinking water consumption data due to differing metering zones, and electricity consumption data include industrial production. A previous study from Norway used mobile phone data from 2016 to train models predicting population based on drinking water production, electricity consumption, and ammonium levels in wastewater. The ammonium model showed the highest accuracy ($R^2 = 0.88$). When tested with 2017 data, all models had mean absolute errors below 10%. The ammonium model estimated Oslo's 2017 population at 645,000, 4% higher than official reports, while drinking water and electricity models significantly misestimated the population (Baz-Lomba et al., 2019). This suggests that future studies should explore multiple data sources, addressing challenges in comparing diverse datasets to refine predictive models.

Improve the data-driven model, the statistical analysis and the approach by including prior engineering knowledge into the analysis

In this feasibility study, the initial linear correlation analysis could benefit from a more advanced statistical approach, possibly incorporating factors like school holidays and catchment characteristics. Exploring non-linear models, such as random forest regression, could also enhance the model. Feature selection, including census and traffic data, offers valuable insights into the mobile phone data and wastewater production relationship. Determining the optimal time resolution for sub-catchments and monitoring sites will probably always be challenging, impacting analysis accuracy. Considering a hydrodynamic model and topography for flow time estimation and accounting for the effects of hydraulic structures, pumping stations, and elevation on wastewater flow could possibly improve correlation accuracy.

Outlook on the future potential of mobile device data in urban water management applications

Operation and planning of wastewater systems

Mobile data has diverse applications to improve the operation and planning of wastewater systems from water pollution control, to distributed estimates of generated wastewater and stormwater volumes as well as pollution and the control of monitoring data quality.

Distributed flows and wastewater pollution: Mobile data aids in estimating sewage flow and water pollution, identifying high pollution areas, and formulating effective pollution control strategies. Mobile phone data enables the estimation of distributed wastewater pollution across various spatio-temporal scales, offering more accurate data than traditional sensor-based methods. This assists in pinpointing pollution sources and developing targeted control strategies. This is also relevant for influent data generation, where dynamic influent data for WWTPs could be generated, enhancing parameterization of influent generators for dynamic simulations and design studies.

Groundwater Infiltration: Mobile data contributes to estimating groundwater infiltration and rain-induced inflow for more accurate wastewater production data. Establishing stable relationships across spatial scales aids in identifying areas where groundwater enters the sewer system, supporting strategies to reduce infiltration.

Stormwater Pollution: Through detailed analysis of mobility patterns, especially traffic, mobile data helps estimate stormwater pollution, particularly from sources like tire wear. This information identifies areas with high pollution levels, enabling targeted pollution control strategies.

Real-time Control Strategies: Mobile data facilitates the prediction and optimization of pollution-based real-time control in urban drainage. Analyzing mobile data allows predicting pollution levels and developing real-time control strategies, integrating rainfall data to determine pollution control measure cut-off levels.

Surveying wastewater generation: Real-time monitoring of significant wastewater producers is another critical use of mobile data in WWTPs. By delineating polygons around specific sites like ski resorts or industrial areas, mobile data helps estimate the volume and composition of non-residential wastewater discharged into the sewer system. This information improves WWTP operation, triggers monitoring campaigns, identifies pollution sources, and informs strategies to mitigate environmental impact.

Population estimates: Mobile data can be used to estimate the population connected to WWTPs. These estimates, derived from mobile phone data, aid in designing and optimizing WWTP performance. Understanding population density and wastewater production fluctuations allows for designing WWTPs to handle anticipated flow rates, enhancing treatment process efficiency. Additionally, it has billing implications for shared WWTP costs among municipalities.

Drinking water applications

Drinking water applications of mobile data seem very promising as well.

Detection of drinking water leaks in distribution systems can be an important use case. Mobile data can be used to monitor the pressure and flow rates within the system and detect any irregularities, in comparison to a reference state, that may indicate a leak. By detecting leaks early, water loss and associated costs can be minimized.

Real-time control of drinking water systems is another application of mobile data. With the help of sensors and mobile data, water utilities can monitor the water quality in real-time and control the treatment processes accordingly. For example, if there is a sudden increase in turbidity, the system can automatically adjust the coagulant dosing to maintain the required water quality standards.

Drinking water demand forecasting can also be improved using mobile data, with the aim to optimize the water supply and treatment processes of a utility. By analyzing the historical data on water consumption patterns, utilities can forecast the demand for water in different locations and at different times of the day, and ensure that the water supply meets the demand.

In addition to municipal water supply, mobile data can also be used to **monitor and optimize industrial water use**. By analyzing mobile phone data from employees or visitors to industrial sites, it is possible to estimate the water usage patterns of individual companies or facilities. Assuming that mobile phone data are available ahead of time than the water use occurs, This information can be used to develop strategies for reducing water consumption, e.g. through dashboards and improving the overall sustainability of industrial operations.

Wastewater-based epidemiology

Since the onset of sanitary engineering, wastewater-based epidemiology (WBE) has proved very useful in the early detection of diseases and the tracking of the prevalence and spread of infectious diseases within a community or region (Sikorski and Levine, 2020; Wilson, 1928). By analyzing wastewater samples over time, researchers can identify trends in disease prevalence, and monitor the effectiveness of public health interventions. As it is cost-effective and provides data in near real-time, WBE is gaining popularity, especially during the COVID-19 pandemic, and has been suggested as a new standard in the proposal of the revised Urban Wastewater Treatment Directive (EC, 2022).

To compare WBE results accurately across time, different locations, and with additional epidemiological data, it is crucial to normalize WBE results to the actual population size. Today, WBE data is commonly normalized with de jure (residential) population numbers. This is not optimal since catchment populations are known to fluctuate (Thomas et al., 2017). The diurnal (e.g. commuters) and intra-annual (e.g. holidays) population fluctuations within WWTP catchment boundaries are often considered to follow a regular pattern (except for special events, e.g. festivals, pandemic measures). This allows the comparison of 24-h loads of respective days over time and the interpretation of variations as changes in consumption/excretion. The evaluation of the validity of this assumption is often difficult and it generally limits the use of the WBE approach to the comparison of "similar" days. The pandemic measures, such as local or regional lockdowns, or working from home, changed the population mobility behavior. This fundamentally challenges normalization with de jure data and emphasizes the need for novel approaches to normalize the de facto population, e.g. with human biomarkers (O'Brien et al., 2014) or mobile phone data. In addition, mobile phone data could also be accompanied by relevant socio-demographic information, such as proportion of males and females or the age distribution of the population.

Potential of wastewater monitoring data to improve mobile phone data

Interestingly, wastewater flow measurements can improve mobile phone data, especially in areas, where the market penetration of a telecom provider is incomplete. This can be through permanent adjustments in catchments with few clients or temporary corrections during network outages. Our case study suggests that wastewater flow data can improve mobile phone data by addressing biases from phone ownership, market penetration, and usage patterns. Telecom providers, by comparing estimates with foul sewage dynamics, can probably understand uncertainties in their data and adjust their data analytics accordingly. Wastewater flows also reveal temporal dynamics in population movement, especially during holidays, complementing mobile phone data limitations. In our view, this can also be particularly useful in small or complex catchments where privacy concerns and cut-off effects may compromise mobile phone data reliability (Figure SI 2), making wastewater-flow derived estimates more accurate than mobile phone data.

Conclusion

With the increasing digitization of society, alternative data sources such as mobile phone data can be used to better assess wastewater production and dynamics than currently possible. In this case study, which includes five diverse catchments in Switzerland, we investigate the relation between mobile phone data and wastewater flows using data from the largest Swiss telecom provider and simple multiple linear regression modelling. Based on our results, we draw the following conclusions:

- Accurately quantifying wastewater flows and pollution loads is a central challenge to effective wastewater management. Current online monitoring technology is not satisfactory due to high costs and, in the case of pollution sensors, limited performance in raw wastewater and sewers applications.
- The mobile phone data from the largest Swiss telecom provider used in this study allows for investigations into the effect of population mobility on wastewater generation. The data were successfully investigated for an entire WWTP catchment as well as for four subcatchments for different temporal aggregations. They were provided at the required spatial resolution and with a temporal resolution of 15min, which is satisfactory for many applications in urban water management. Unfortunately, we found unexpected phenomena in the data, e.g. drop of population estimates around midnight, which we could not fully explain. Also, unknown biases due to service provider coverage remain a challenge and privacy-related data processing issues inhibit the use of mobile data on small spatial and temporal scales.
- We successfully applied a simple multivariate linear regression approach for time series which could be improved by incorporating past population counts to account for travel times in the sewer.
- The results demonstrate the potential of using mobile phone-based population density data to predict wastewater flows, particularly for smaller sub-catchments, because spatio-temporal changes in population are averaged out stronger in larger catchments.
- Despite promising results, it is clear that the relationship between mobile data-based population and wastewater flow is non-linear, which calls for corresponding modeling approaches to improve flow predictions. One possibility to enhance prediction could be to incorporating prior knowledge on the sewer systems, e.g. through catchment specific flow times derived from sewer system topography or even a rainfall-runoff model. Also, further alternative data sources (e.g. traffic data) could improve predictions.
- Mobile phone data and the population dynamics they reflect, present significant potential not just in urban drainage contexts but also in other urban water management applications, including the assessment of industrial wastewater production, the optimization of drinking water networks and the population normalization in wastewater-based epidemiology.

- The research underscores the relevance of digitalization and open data for urban water management. In the future, further non-traditional data sources, such as mobile phone traces or traffic data, can contribute to predictive capabilities in urban water management applications, potentially leading to more effective and efficient wastewater generation prediction and pollution control strategies. Important aspects are data protection and privacy issues, which might call for domain-specific data processing and sharing.

Acknowledgements

We would like to express our sincere gratitude to all individuals who contributed to the completion of this paper. Their valuable support and assistance were instrumental in shaping our research and enhancing the quality of our work. Special gratitude to the colleagues from Abwasserverband Altenrhein, Frank Lükewille and Christoph Egli, who helped greatly in selecting suitable subcatchments, provided the sewer monitoring data and metadata. We thank the Swisscom employees who have been supporting our use case over the years, namely Andrea Schwaller, Yann Steimer and Leandro Kieliger and Res Witschi. We deeply appreciate the contributions of each individual mentioned above, as well as anyone else who has contributed to this paper indirectly. Without their assistance, this study would not have been possible.

Author contribution statement

The research was conceptualized and designed by JR and AD. Data collection and analysis were conducted by JR, AD and NN. The manuscript was written and revised by JR, AD, NN, SB and CO, who approved the final version for submission.

Availability of data and code and license for re-use

Data from this study are available at the Eawag Research Data Institutional Repository (https://doi.org/10.25678/000D2B). The dataset "Mobile Device Data for Urban Water Management: Predicting Wastewater Flows" © 2023 by Jörg Rieckermann, Andy Disch, Eawag, AV Altenrhein is licensed under CC BY-SA 4.0. To view a copy of this license, visit http://creativecommons.org/licenses/by-sa/4.0/
The current version of the code is available on https://zenodo.org/records/12958776 and will be integrated into https://gitlab.switch.ch/sww/mobile_flow_data_altenrhein_2019.

References

Agustsson, J., Akermann, O., Barry, D. A., and Rossi, L.: Non-contact assessment of COD and turbidity concentrations in water using diffuse reflectance UV-Vis spectroscopy, Environ. Sci. Process. Impacts, 16, 1897, https://doi.org/10.1039/C3EM00707C, 2014.

Atinkpahoun, C. N. H., Le, N. D., Pontvianne, S., Poirot, H., Leclerc, J.-P., Pons, M.-N., and Soclo, H. H.: Population mobility and urban wastewater dynamics, Sci. Total Environ., 622–623, 1431–1437, https://doi.org/10.1016/j.scitotenv.2017.12.087, 2018.

AVA: Abwasserverband Altenrhein Geschäftsbericht 2019, 2020.

Baz-Lomba, J. A., Di Ruscio, F., Amador, A., Reid, M., and Thomas, K. V.: Assessing Alternative Population Size Proxies in a Wastewater Catchment Area Using Mobile Device Data, Environ. Sci. Technol., 53, 1994–2001, https://doi.org/10.1021/acs.est.8b05389, 2019.

Bisinella de Faria, A. B., Besson Mathilde, Ahmadi Aras, Udert Kai M., and Spérandio Mathieu: Dynamic Influent Generator for Alternative Wastewater Management with Urine Source Separation, J. Sustain. Water Built Environ., 6, 04020001, https://doi.org/10.1061/JSWBAY.0000904, 2020.

Bourgeois, W., Burgess, J. E., and Stuetz, R. M.: On-line monitoring of wastewater quality: a review, J. Chem. Technol. Biotechnol., 76, 337–348, https://doi.org/10.1002/jctb.393, 2001.

Brokamp, C., Beck, A. F., Muglia, L., and Ryan, P.: Combined sewer overflow events and childhood emergency department visits: A case-crossover study, Sci. Total Environ., 607–608, 1180–1187, https://doi.org/10.1016/j.scitotenv.2017.07.104, 2017.

van Daal-Rombouts, P., Schilperoort, R., Langeveld, J. G., and Clemens, F.: CSO pollution analysis based on conductivity and turbidity measurements and implications for application of RTC, in: Novatech 2013 - 8ème Conférence internationale sur les techniques et stratégies durables pour la gestion des eaux urbaines par temps de pluie / 8th International Conference on planning and technologies for sustainable management of Water in the City, Lyon, France, 2013.

EC: Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL concerning urban wastewater treatment (recast), 2022.

Gruber, G., Winkler, S., and Pressl, A.: Continuous monitoring in sewer networks an approach for quantification of pollution loads from CSOs into surface water bodies, Water Sci. Technol., 52, 215–223, https://doi.org/10.2166/wst.2005.0466, 2005.

Gruber, G., Pichler, M., Hofer, T., Maier, R., and Clara, M.: Messtechnische Ermittlung von Jahresschmutzfrachten in einem Mischwasser- und einem Niederschlagswasserkanal, in: Tagungsband -

Aqua Urbanica 2019: Regenwasser weiterdenken - Bemessen trifft Gestalten, Aqua Urbanica 2019: Regenwasser weiterdenken – Bemessen trifft Gestalten, 107–118, 2019.

Lepot, M., Torres, A., Hofer, T., Caradot, N., Gruber, G., Aubin, J.-B., and Bertrand-Krajewski, J.-L.: Calibration of UV/Vis spectrophotometers: A review and comparison of different methods to estimate TSS and total and dissolved COD concentrations in sewers, WWTPs and rivers, Water Res., 101, 519–534, https://doi.org/10.1016/j.watres.2016.05.070, 2016.

Li, C., Sylvestre, É., Julian, T., and Kohn, T.: Risk assessment of waterborne virus in Lake Geneva: the present and the future, EGU-4946, https://doi.org/10.5194/egusphere-egu23-4946, 2023.

McGinnis, S. M., Burch, T., and Murphy, H. M.: Assessing the risk of acute gastrointestinal illness (AGI) acquired through recreational exposure to combined sewer overflow-impacted waters in Philadelphia: A quantitative microbial risk assessment, Microb. Risk Anal., 20, 100189, https://doi.org/10.1016/j.mran.2021.100189, 2022.

O'Brien, J. W., Thai, P. K., Eaglesham, G., Ort, C., Scheidegger, A., Carter, S., Lai, F. Y., and Mueller, J. F.: A Model to Estimate the Population Contributing to the Wastewater Using Samples Collected on Census Day, Environ. Sci. Technol., 48, 517–525, https://doi.org/10.1021/es403251g, 2014.

O'Brien, J. W., Thai, P. K., and Tscharke, B. J.: Population biomarkers for wastewater-based epidemiology, 123–138 pp., https://doi.org/10.1016/B978-0-443-19172-5.00005-6, 2023.

Ort, C.: Quality assurance/quality control in wastewater sampling, in: Quality Assurance & Quality Control of Environmental Field Sampling, Future Science Ltd, 146–168, https://doi.org/10.4155/ebo.13.476, 2014.

Palmer, J. R. B., Espenshade, T. J., Bartumeus, F., Chung, C. Y., Ozgencil, N. E., and Li, K.: New Approaches to Human Mobility: Using Mobile Phones for Demographic Research, Demography, 50, 10.1007/s13524-012-0175-z, https://doi.org/10.1007/s13524-012-0175-z, 2013.

Pistocchi, A. and Dorati, C.: Combined Sewer Overflow Management: Proof-of-Concept of a Screening Level Model for Regional Scale Appraisal of Measures, in: springerprofessional.de, UDM2018, Palermo, Italy, 2018.

Sikorski, M. J. and Levine, M. M.: Reviving the "Moore Swab": a Classic Environmental Surveillance Tool Involving Filtration of Flowing Surface Water and Sewage Water To Recover Typhoidal Salmonella Bacteria, Appl. Environ. Microbiol., 86, e00060-20, https://doi.org/10.1128/AEM.00060-20, 2020.

Thomas, K. V., Amador, A., Baz-Lomba, J. A., and Reid, M.: Use of Mobile Device Data To Better Estimate Dynamic Population Size for Wastewater-Based Epidemiology, Environ. Sci. Technol., 51, 11363–11370, https://doi.org/10.1021/acs.est.7b02538, 2017.

Troxler, P.: Predicting Inflow to WWTP Oberengadin - Adaptive Kalman Filter models to improve receiving water protection, Master's thesis, ETH Zurich, 2018.

Wilson, W. J.: ISOLATION OF B. TYPHOSUS FROM SEWAGE AND SHELLFISH, Br. Med. J., 1, 1061–1062, https://doi.org/10.1136/bmj.1.3520.1061-a, 1928.