Mobility data for reduced uncertainties in model-based WWTP design

Oscar Samuelsson , Erik U. Lindblom , Kenneth Djupsjö ,
Linda Kanders , Lluís Corominas

PII: S2589-9147(25)00117-3
DOI: https://doi.org/10.1016/j.wroa.2025.100418
Reference: WROA 100418

To appear in: *Water Research X*

Received date: 16 June 2025
Revised date: 18 August 2025
Accepted date: 20 September 2025

Please cite this article as: Oscar Samuelsson , Erik U. Lindblom , Kenneth Djupsjö , Linda Kanders , Lluís Corominas , Mobility data for reduced uncertainties in model-based WWTP design, *Water Research X* (2025), doi: https://doi.org/10.1016/j.wroa.2025.100418

Highlights

- Mobility data from mobile phones can provide dynamic population estimates
- Seasonal and weekly population variations were easy to quantify
- Mobility data enable dynamic load time series for model applications
- Person loads with lowered variance reduced the uncertainty in model-based design
- Nitrogen and phosphorous loads showed stronger correlation with population than BOD

Mobility data for reduced uncertainties in model-based WWTP design

Oscar Samuelsson[1], Erik U. Lindblom[1,2], Kenneth Djupsjö[3], Linda Kanders[1] and Lluís Corominas[4]

[1] IVL Swedish Environmental Research Institute, Valhallavägen 81, Stockholm, 114 28, Sweden.

[2] Lund University, Division of Industrial Electrical Engineering and Automation (IEA), Department of Biomedical Engineering, Lund University, P.O. Box 118, SE-22100 Lund, Sweden.

[3] Telia Company, Customized Delivery & LCM, Helsinki, Finland.

[4] ICRA Catalan Institute for Water Research, Spain.

Keywords: Design guideline; Dimensioning; Sizing; Wastewater treatment; Water resource recovery facility.

Abstract

Model-based design is an emerging tool for dealing with the uncertain dynamic loads entering wastewater treatment plants (WWTPs). But our understanding about the load-driving population-dynamics is limited. Therefore, we studied if mobility data (mobile telecommunications data) could be used to reduce uncertainties during design. Mobility data from Uppsala, Sweden between 2019–2022 clearly quantified population movement patterns that were useful for simulating load scenarios such as seasonal load-shifts, without data gaps from irregular influent sampling. Further, they showed fair correlations with the daily influent nitrogen load (R^2 = 0.49), which resulted in a more precise person load estimate than assuming a static population (23 % reduced variance). Unfortunately, BOD load variations showed little correlation with the population variations (R^2 = 0.21). Nevertheless, model-

based reactor sizing based on mobility data successfully reduced the de-/nitrification volume safety factor with 5 %, which demonstrates their practical usefulness for WWTP design.

1 Introduction

Rapid urbanization pushes many wastewater treatment plants to retrofit and upsize their capacity. A key challenge is then to safeguard the treatment efficiency, despite the uncertainty in future influent load variations.

The typical engineering practice is to set the future design load by the following actions:

1) Identify the current load per person by normalizing the total load from the population, with the population size.
2) Predict the future population at the design year.
3) Multiply the person load from 1) with the anticipated population in 2).
4) Add anticipated industry load and other non-population related sources to 3).

These steps can be described in mathematical terms (Section 4.2) with a statistical data model.

A challenge in the design load computation is to obtain a good person load estimate in 1), since day-to-day population variations are unknown. Commonly, a constant population indicated by yearly population records is used as a proxy, which only accounts for resident persons that are registered within the sewer network area. This simplification may bias the person load, and effectively transfer any bias to the total design load in 3).

A second challenge is to separate non-population related loads from being lumped with the total load in step 1). Unless industry and other non-population related loads (henceforth referred to as base load), are subtracted from total load, the person load will be overestimated and bias the design load.

Today, state-of-the-art design leans on dynamic modelling and simulation for making informed decisions (Belia et al., 2021; Yang et al., 2022). But models are data hungry, calling for dynamic influent load time series of BOD, COD, phosphorous, and nitrogen components. Although on-line sensors exist for these parameters, they are seldom used in the influent due to the harsh conditions that require extensive sensor maintenance. For design purposes, historic load data are therefore mostly limited to 24-h composite samples. This leads to a

dynamic-data gap where load variations are assessed with irregular sampling, with, at the best, daily averages a few days a week.

This data gap has triggered the development of so-called influent generators, which can produce any imagined influent load and flow – with complete data, see e.g. (Gernaey et al., 2011; Li and Vanrolleghem, 2022; Talebizadeh et al., 2016). But the current influent generators have been developed with different purposes in mind. Benchmarking control strategies where the early focus in (Gernaey et al., 2011), where seasonal and industry variations were possible to manually tweak to test different load scenarios. Data-driven approaches have also been used (Li and Vanrolleghem, 2022) where historic data can be regenerated with varying load. Concentration dilution can be simulated by focusing on the flow rate variations (Talebizadeh et al., 2016). More recently, model-based influent estimation methods have been proposed (Alex, J., 2024; Wärff et. al., 2025), as a solution to the real-time data needs for digital twins.

However, during design, the future load is predicted from a population prognosis. To our knowledge, influent generators so far does not consider population as an underlying driver for the influent load, although this could reduce the uncertainties during extrapolation and estimation of the future design load. Both on a short timescale (minutes to hours, e.g., for aeration system design), and for predicting seasonal variability such as weekend-to-weekday shifts in tourist attractive areas. Altogether, model-based design might benefit from an improved understanding of population variations, and how they are reflected in the influent load.

One opportunity is to use data from the communication between mobile phones and base stations to estimate and analyse dynamic population variations. Such data, henceforth referred to as mobility data, have been used to normalize micropollutants and pathogens in the water based epidemiology domain (Baz-Lomba et al., 2019; Gudra et al., 2022; Sim et al., 2023). However, they have not yet been considered in a WWTP design context. The objective of this paper is therefore to explore how mobility data can support model-based design.

2 Results and Discussion

2.1 Mapping mobility data with the sewer catchment area

The common approach for estimating the population in step 1) is to identify the houses and properties within the sewer's geographic area, which, in practice is defined by a specified area in the geographic information system (GIS). Then, the registered residents within this area (typically on a yearly basis) indicate the population, henceforth referred to as static population estimates.

The alternative data source explored in this study are mobility data. These are produced from the network communication between mobile phones and antennas in a well-defined area (Section 4.1.2). In short, mobility data provides a high frequency estimate (down to 20 min) of mobile active persons within squared GIS-areas.

However, these squares don't align with the irregularly shaped sewer catchment area, which makes it difficult to match mobility data with the sewer area. Therefore, a normalization factor was developed to exclude the population, which was registered within the mobility data squares, but is outside the sewer area (Figure 5 in Section Materials and Methods). This normalization was obtained as the ratio between the static population within the sewer area, and within the larger mobility data squares (Section 4.1). The static population was used since it is available for any area shape. The normalization factor obtained for 2019 was 96 %, which indicates that 96 % of the static population within the mobility data squares, also were residents within the sewer catchment area . This normalizing method was found to be more representative than using the actual land area as normalizer, since the rural land outside the sewer area was unpopulated (Figure 5).

2.2 Mobility data reflect population habit dynamics

The difference between the static population estimates and the population indicated by mobility data is evident (Figure 1a), where mobility data shows seasonal dynamics that are related to big holidays and the vacation period in June–August. Part of the drop in population between June and September is caused by the university summer break with 26 000 full-time students (Uppsala universitet, 2023), as compared to the static population estimate of about 185 000 people (Statistics Sweden, 2020). Note the pronounced dip in mid-June, indicated as

"midsummer", which is a Swedish holiday that is traditionally celebrated on the countryside, hence outside the city.

The clear seasonal variations suggest several interesting load-shifts to simulate and handle in the design phase. For example, the return from school break in mid-February could represent (a likely) extreme scenario, with the "extreme" being a cold wet-weather event coinciding with the instant increase in load from 150 000 to 180 000 persons. Similarly, the substantial population variations during Easter and Christmas quantify likely load-shifts that could be planned for in the operational strategy during design, much like a long-term feed-forward controller.

From a design perspective, the week-to-week variation is also important. In particular, the maximum average weekly load is specified in the permit for Swedish WRRFs and should be compared with the average weekly design load. Figure 1(c) shows the non-Gaussian shape of the weekly mean population where standard working weeks with about 180 000 people clearly stand out as the mode, which may be useful for simulating permit compliance.

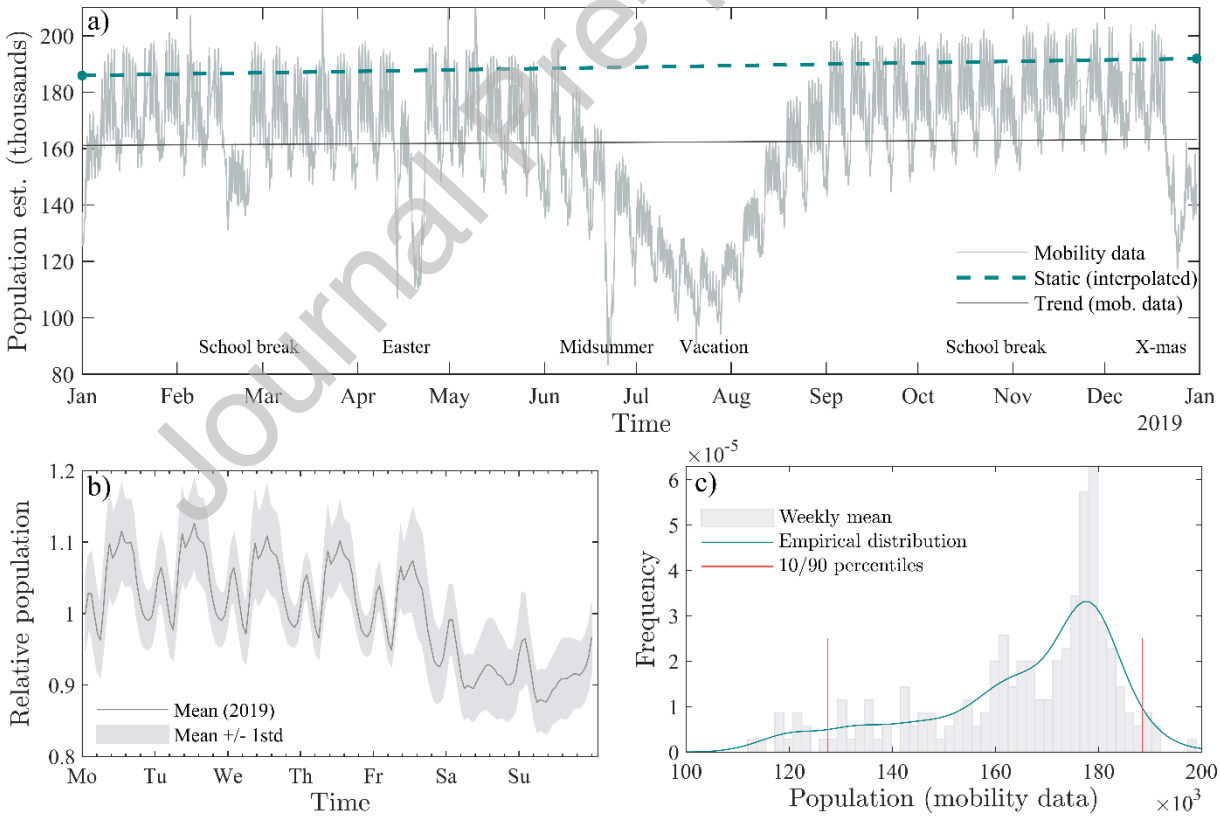


Figure 1. (a) Difference between mobility data (grey solid line) and static population estimates (green dashed line) within Uppsala sewer catchment area 2019. (b) Population variation during weeks, when normalized to the weekly

average population. The small peaks every night are anomalies caused by the anonymization step described in Section 4.1.2. (c) Empirical probability density estimates of the weekly mean population (green solid line) with 10 and 90 percentiles indicated with red vertical lines.

Looking back on the details in Figure 1(a), there is a persistent noise in mobility data. A closer look reveals that the noise is in fact periodic and caused by weekly population variations (Figure 1b). The normalized population data indicates a 10 percent reduction during weekends and the opposite increase during weekdays (Figure 1b). A reduced wastewater generation during holidays has been found to indicate a high socio-economic area (Corominas et al., 2024), but here also non-resident commuters are likely also contributing to the differences between weekdays and holidays. Note that Figure 1 is based on data during 2019, and the norm before the pandemic 2020–2022. These weekly patterns may be useful for simulating effects from population behaviour, and their impact on load dynamics. Specifically, the increase in hybrid work and related changes in population movements can't be analysed from historic data but needs to be simulated based on socioeconomic and behavioural prognoses.

2.3 The impact of population estimates as a design load normaliser

As mentioned in Introduction, an accurate person load is key in design where the total load is measured in terms of influent biological oxygen demand (BOD) and nitrogen mass per day. The daily load is then normalised by the estimated population and finally presented as an annual mean person load.

Figure 1(a) shows that there are clear differences between the population estimates with, on average, 25 000 more people in the static population estimate as compared to the mobility data. This difference is, in part, explained by the fact that the mobility data model deliberately excludes children below 6 years (simply because they don't use mobile phones). These count to 7 400 children for the whole Uppsala municipality during 2019 (Statistics Sweden, 2024). The remaining population difference is caused by the seen population decrease in mobility data during summer.

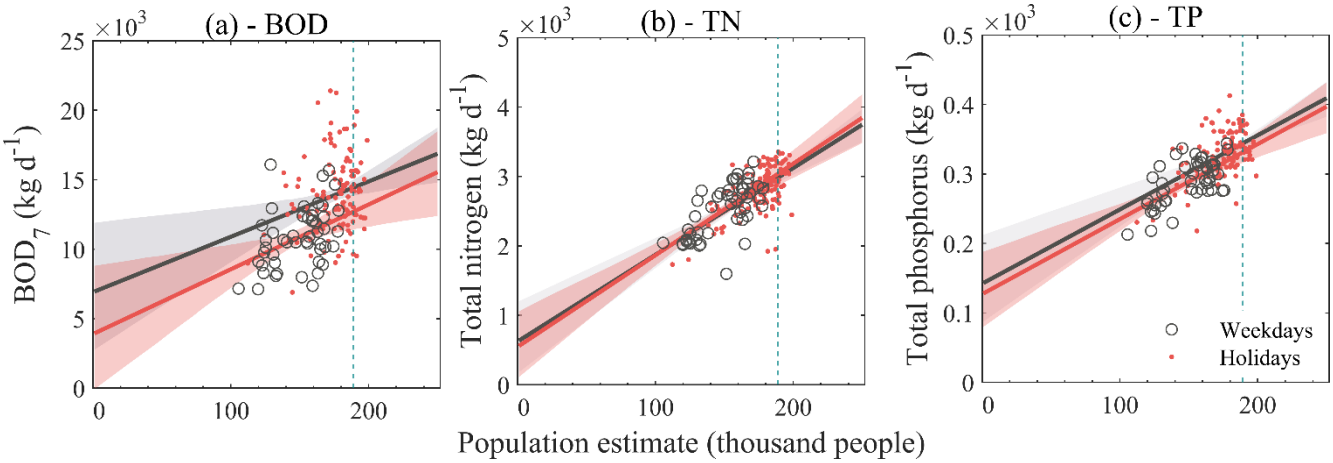
Logically, the different population estimates resulted in differing person load estimates with 68 g and 78 g BOD per person and day for the static population and mobility data, respectively.

At first, one may argue that such 15 percent person load difference indicates an equivalent difference in predicted design load (and required plant capacity). However, the population estimate impact on design load is likely less influential if the same population source is used for both normalization and prognosis. That is, the static population may be an overestimate compared to mobility data (causing a low person load) but using the same data source it would likely also suggest a larger population increase, and thereby end up in a similar design load as mobility data, in the end. With such reasoning, we suggest that it may be important to stick with one population source throughout the design process, regardless of whether this is mobility data or a static measure.

2.4 Is population a good predictor for wastewater load?

However, so far, we have not studied whether mobility data really are good predictors for the wastewater load. Therefore, we now analyse the correlation between the daily population and corresponding BOD, total nitrogen (nitrogen), and total phosphorus (phosphorous) loads. If the load data follow the model with a person load, we expect a straight-line relationship with respect to population variations (Equation 1 in Section 4.2).

Unfortunately, the BOD load show little correlation ($R^2 = 0.21$) with the population (Figure 2a) and is difficult to predict from mobility data. Similarly, total organic carbon (TOC) had a poor correlation ($R^2 = 0.08$, data not shown). However, nitrogen, and phosphorus showed fair correlations with the estimated population (Figure 2b and 2c) with R^2 values of 0.49. This is in line with (Baz-Lomba et al., 2019) that concluded ammonia to be the best population predictor (as compared to electricity and drinking water), when validated with mobility data. Further, nitrogen and phosphorus loads show a straight-line relationship with the population, which supports the data model with a constant person load (Section 4.2).



The BOD7 on y-axis, in (a) will be replaced with BOD in the final figure.

Figure 2. Linear regression fits for estimating the person load (slope) and baseload (intercept) from daily influent load data of BOD (a), total nitrogen – TN (b), and total phosphorus – TP (c); versus mobility data. Static population estimates are indicated with vertical green dashed lines. Black symbols refer to weekday and red holidays. Solid lines show best fit regression lines with shaded areas for 95 % bootstrap confidence intervals.

But for design, a good nitrogen prediction is not sufficient when BOD is unpredictable since unexplainable noise in BOD also impact the BOD-to-nitrogen ratio. This ratio is critical and decides the required pre-denitrification volume and needed dosage of external carbon source (Lindblom and Samuelsson, 2023).

Therefore, we reconsidered the BOD load data and noticed that the average BOD person load differed depending on weekday (*t*-test at 0.05 significance level). In specific, both the person load and its variability were lower during holidays, as compared to weekdays (Table 1, Method A –Difference in means). Likewise, the correlation coefficients between BOD and nitrogen showed a stronger correlation on holiday data (mobility data: 0.40, static: 0.57), than on weekdays (mobility data: 0.1, static: 0.24). It should be noted that nitrogen showed similar variance and person load values, regardless of weekday.

To conclude, mobility data best predicts nitrogen load, regardless of weekday, whereas BOD is difficult to predict, albeit with slightly lower variance during weekends.

2.5 Can we blame industry for the BOD load variability?

But why is the BOD person load variability lower during holidays? A natural guess is that the industry load (typically attributed with BOD/COD streams) is smaller during holidays. If so, one solution would be to estimate the base load distribution, and thereafter subtract it from the

total load, and potentially also reduce the unexplainable weekday variability. We tried two methods (Section 4.2.2) to estimate the base load as:

A) The added load during weekdays by subtracting the population load, which, in turn, was estimated from holiday data.
B) The intercept from a linear regression of the population variations and load, which indicates zero persons (Figure 2).

These were compared with the current approach that uses emission reports from the largest industries.

The results are given in Table 1 where method A shows a BOD base load in line with the industry emission data. Note that negative base loads are indicated for other components than BOD, which is an effect of having a lower person load during weekdays as compared to holidays. In fact, this result supports the previous observation that nitrogen person load is similar regardless of weekday and possibly not affected by a base load.

Method B indicates a six times higher base load than method A. Additionally, the uncertainty in the regression is high (see confidence intervals in Figure 2(a) and Table 1). This is an effect of the low correlation between BOD load and mobility data. Thus, method A seems to be the better approach for estimating the baseload, although not necessarily better than the current method based on emission reports. In fact, reconsidering the noisy BOD load data one could question whether the general assumption with a person specific BOD load is valid. Our conclusion is that more data, from other data sources, are needed to understand the BOD variability, and to predict its future load.

Table 1. Statistics for differences between holiday and weekday specific person loads 2019–2022. Estimated BOD baseloads are emphasized in bold. Acronyms are biological oxygen demand (BOD), total nitrogen (TN), total phosphorus (TP), and total organic carbon (TOC).

		Method A) – Difference in means					**Method B) – Linear regression**							**Reference – Emission data**
		n	***Population***		***Baseload***		***Population*** *(slope)*			***Baseload*** *(intercept)*			***R^2***	***Industry load***
	Unit	*#*	(g p^{-1} d^{-1})		(kg d^{-1})		(g p^{-1} d^{-1})			(kg d^{-1})				(kg d^{-1})
			mean	*SD*	*Mean*	*SD*	*mean*	*2.5 %*	*97.5 %*	*mean*	*2.5 %*	*97.5 %*		2018/2019; 2020/2021/2022
BOD	Weekday	121	79.7	15.7	**1 070**	217	39.8	10.0	69.6	**6 910**	1 670	12 200	0.06	**940/440; 450/1200/1099**
	Holiday	49	73.1	13.9			46.6	16.4	76.8	3 900	- 652	8 450	0.17	
TN	Weekday	122	16.1	1.57	- 136	*n.a.*	12.9	9.86	16.0	554	13.3	1 090	0.37	41/48; 54/180/184
	Holiday	49	16.9	1.89			13.2	8.92	17.5	546	- 98.2	1 190	0.45	
TP	Weekday	121	1.89	0.184	- 11	*n.a.*	1.11	0.781	1.44	136	78.0	193	0.27	20/21; 19/31/33
	Holiday	49	1.94	0.196			1.08	0.715	1.45	127	72.0	182	0.43	
TOC	Weekday	122	60.6	20.9	- 247	*n.a.*	40.3	- 0.355	81.0	3 530	- 3 620	10 700	0.031	*n.a.*
	Holiday	48	61.5	22.1			56.3	4.05	109	752	- 7 090	8 600	0.093	

2.6 Can mobility data support model-based plant design?

One aspect that we have overlooked so far is the role of sampling time during design. This is important since some process units, e.g., the aeration system, are designed for hourly or minute peak-loads, by contrasts to bioreactor volumes that are designed for monthly or yearly average loads.

Figure 3 shows three dry-weather load profiles measured with bihourly resolution next to mobility data. Clearly, the correlation is low at this resolution ($R^2 < 0.1$), where the load profiles show more pronounced dynamics than the population estimates. This is reasonable since our food intake (and waste secretion) follows a fixed pattern that is not evenly distributed throughout the day and night.

But dry weather profiles are essential for dynamic simulations and model-based design. A future research study is therefore needed to analyse how the daily profile change with an increasing/decreasing population, and how this could be modelled to produce high resolution design scenarios.

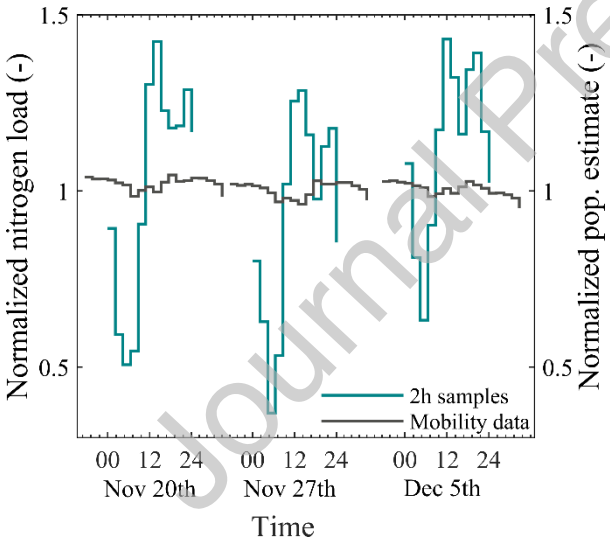


Figure 3. Bihourly data for daily nitrogen load profiles and mobility data for three dry-weather days, when normalized to daily mean values.

On the other end of the timescale, seasonal load variations need to be analysed. But the dynamic data gap with incomplete historic time series (due to irregular influent sampling) limits whole-year simulations. Here, mobility data can support by generating time series that fill the dynamic data gap. Figure 4(a) shows how mobility data have been multiplied with the

estimated person loads to generate a complete load time series. Such time series are practical in model-based design when simulating the permit compliance for a whole year. Figure 4(a) shows complete time series for both the present population and for the design year 2050 with an assumed population increase to 330 000 people (that follow the same weekly patterns as today, see Section 4.3). Any altered population-related scenarios, e.g., trends in person load values or commuter patterns could be generated straightforwardly.

Finally, we analyse how the choice of population estimates, via person load estimates, impact a model-based WWTP design. We use the method and conventional plant configuration in (Lindblom and Samuelsson, 2025) to size de-/nitrification volumes with the design population 2050 in Figure 4(a) as input. Since the person loads were estimated as a probability distribution (Section 4.2.3), it is straightforward to generate random samples of person loads and propagate them through the model-based design as an uncertainty analysis.

Figure 4(b) shows the required reactor sizes (dots) from 1 000 Monte Carlo simulations, with an ellipsoid indicating the 60th percentile of an empirical kernel density estimate (the 60th influent flow percentile is a common design choice in Sweden). As expected, the most likely design (circle) increases with increasing person load (mobility data showed a 15 percent larger person loads in Section 2.3). Further, the higher variance in the static population estimates resulted in a greater design uncertainty. In fact, the needed safety factor for the total volume (here defined as the difference between the ellipsoid mode and the maximum volume in the 60th percentile) was 5 percent larger for static data, as compared to mobility data (19 % versus 13 %). This demonstrates the usefulness of mobility data for model-based design.

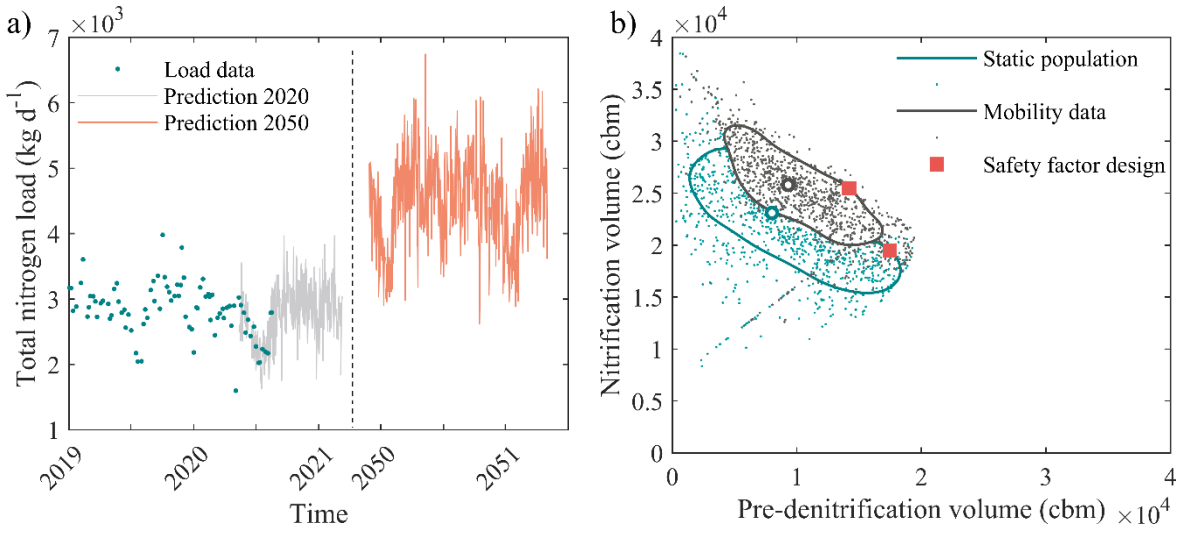


Figure 4. (a) Example of how mobility data can be used to generate complete influent load time series for current load (grey line) and a predicted design load (red line). (b) Difference in reactor volumes for model-based design of the predicted population in (a) when using static population estimates or mobility data for estimating person loads. The dots represent the 1 000 Monte Carlo-simulations, each representing one design and the corresponding Gaussian kernel density estimate (60th percentile ellipsoids), most likely design (the mode, circle), and maximum 60th percentile volume considered as the safety factor design (red squares).

3 Conclusion

Mobility data could support model-based design by:

- Reducing the variance in person load estimates, which, in effect, reduced uncertainties in the reactor safety factor by 5 %.
- Quantifying seasonal population variations that can be translated and extrapolated to whole-year simulation scenarios for design.

These findings demonstrate that data from outside-the-fence are available and useful for understanding the population dynamics and the related dynamic influent load.

This study emphasized the importance of having access to complete dynamic load time series for reducing uncertainties in model-based design. However, high quality time series describing seasonal and daily variations are of general interest in model applications, including the trending digital twins. We therefore suggest more research on mobility data and other socioeconomic data sources that relate to the WWTP influent. These data may complement online sensor and laboratory data, in particular for explaining BOD load variations.

4 Materials and Methods

4.1 Data

In this paper we explore a new data source, mobility data, and evaluate its usefulness for WRRF design by comparing it with the more common static population record. The static population records are explained in Section 4.1.1, and the mobility data are described in Section 4.1.2. Apart from this new data source, only standard influent data were used to characterize the wastewater, which are described in Section 4.1.3.

4.1.1 Static population records

The state-of-the-art approach for estimating the population is to first identify the properties and houses within the sewer's geographical area, typically via the geographical information system (GIS). Then, the registered residents make out the population estimate. Such population estimates are typically obtained on a yearly basis from resident census data, and slowly change due to the net population change. Here, we refer to this as the static population estimate.

In this study, the sewer GIS-area for Uppsala, Sweden, was matched with the residents registered by the governmental agency that is handling population statistics, Statistics Sweden (in Swedish: Statistiska centralbyrån – SCB). The sewer area included the main city alongside with a few surrounding villages (Figure 5) that were connected to the main plant Kungsängsverket.

4.1.2 Dynamic population estimates – mobility data

An alternative data source, which is explored in this study are mobility data. These data are produced from the network communication between cell phones and antennas in a well-defined area. A short description of the mobility data is provided here, and details can be found on the data supplier's website (Telia, Crowd insights).

Raw cellular data from the network communication between cellular devices and antennas are converted to so-called mobility signals. The mobility signals contain geographical information about the cellular device location. However, the mobility signals cannot be used directly, but first needs to be anonymised.

In short, all cellular devices are provided random identifiers to separate information about the device, from its positional data. Also, the identifier is replaced at a 24-hour interval, making it impossible to extract time-series for consecutive days. To further ensure anonymity in the mobility data, only groups of 5 or more people are considered per geographic area and temporal resolution. The sampling time is available down to 20-minutes, although we here used 24-h and 2-h measurements to match the corresponding wastewater sampling rates.

Only the total number of devices within a certain GIS-area (a square within GIS system) are provided as mobility data. That is, the exact position of the different devices cannot be obtained, but only its location up to a certain square. Further, the square size ranges between 250 m × 250 m and 1 000 m × 1 000 m to assure a minimum number of devices within each square for integrity reasons (Figure 5).

The sewer catchment area is however not square (see red areas in Figure 5). To only account for the population within the sewer catchment area, two simplifications were made.

1) Any square containing a part of the sewer area was included as part of the whole "sewer population". This produced a large area built-up by squares denoted *the complete population area*.
2) A normalization factor was applied to correct for households within the complete population area but, which were outside the sewer catchment area (mainly rural areas as indicated in Figure 5). The normalization factor was obtained as the percentage of the static population within the sewer area, as compared to the static population within the complete population area. During 2019, this percentage was 96 percent, which was used to normalize the mobility data.

The model translating cellular network data to a population estimate was developed by the data supplier. Hence, from our perspective, the data were regarded as black-box output from a dynamic population sensor. A key part of the black-box model is to extrapolate the data supplier's 35 percent market share, to the whole population. It should be noted that the whole population in the model implicitly excludes people without a cellular phone. In Sweden, few children below 6 years of age carries a cellular phone and thus the dynamic population is underestimated. This is a deliberate decision by the data supplier, not trying to compensate for this young population proportion.

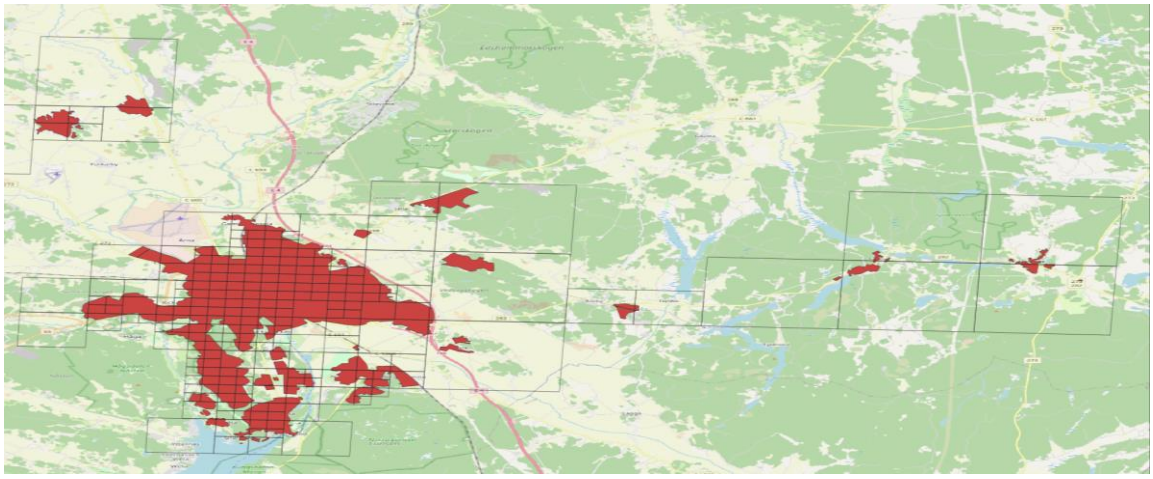


Figure 5. Overlap between the sewer catchment area (red) and the mobility data squares (thin black lines) in Uppsala, Sweden. Note that the square size is smaller in the dense city areas as compared to the rural areas with less people.

4.1.3 Wastewater data

The wastewater was characterized with the influent routine samples consisting of flow (Q_{in}), total nitrogen (TN), and total phosphorus (TP), organic matter measured as biological oxygen consumption during seven days (BOD) and as total organic carbon (TOC). Flow-proportional wastewater samples were collected once a week at the WRRF influent and corresponding industry wastewater loads were obtained annually via emission reports. Four years of load data were considered during 2019–2022. This means that the first year, 2019, was before the pandemic, and the following three years were affected by restrictions such as reduced commuting and tourists.

In addition to the standard data, data from a high frequency sampling campaign (2-h samples during four non-consecutive dry-weather days) were used to assess the short term dynamics in mobility data.

4.2 Load estimates

During design, several assumptions are made regarding the influent load generation. This section details our load assumptions explicitly in a data model, which was used to analyse the load and population data.

4.2.1 Population person load

The person load for a component *x* was defined as the average load per person and day (PL). This person load was analysed during weekdays and weekends. Weekends also include big

holidays, school breaks and the Swedish standard four-week industry vacation period. In total, the weekends amount to about 30 percent of a full year.

4.2.2 Industry and baseload

Three approaches were assessed to separate the population load from the remaining industry and baseload, henceforth denoted method A, B, and the third reference method based on industry emission report data. Method A and B used mobility data.

Method A assume that the industry load is only active during weekdays. Then, an estimate of the person load during holidays can be used to subtract the person load during weekdays, where the remainder is assumed to be the industry load. An implicit assumption here is that the person load is constant, regardless of the day of the week.

Method B uses linear regression with the total load on the *y*-axis and the population on the *x*-axis. The regression line load at zero population then becomes an estimate of the base load. Similarly, the regression line slope estimates the person load. As an example, the regression for BOD would follow the line

	$Q_{in}BOD_{c,in} = BOD_{m,base} + BOD_{PL}\ pop$	(1)

where BOD_{PL} is the BOD person load, $BOD_{m,base}$ is the BOD base load, $BOD_{c,in}$ is the influent BOD concentration, and pop is the population. Note that the base load is likely larger than the industry load since any source contributing to a fixed base load will be included in the base load.

The reference method is the one in current use at the WRRF. The method relies on annual emission reports from the five largest industries. Based on these data, the load during weekdays was estimated as the proportional average of working days (weekdays) for one year. Thus, the reference method underestimates the total base load since it only considers the industry proportion.

4.2.3 Stochastic load model

The total influent load of a component x at time t day was described with the data model

	$x_{tot}(t) = x_{base}(t) + pop(t)\ x_{PL}(t)$	(2a)

	$x_{base} \sim \Phi(m_{x_{base}}, R_{x_{base}})$	(b)
	$x_{PL} \sim N(m_{x_{PL}}, R_{x_{PL}})$	(c)

where the base load x_{base} was assumed to have first and second order moments $m_{x_{base}}$ and $R_{x_{base}}$, respectively, and follow an unknown distribution Φ. The person load x_{PL} was assumed to be Gaussian with mean m_{PL}, and variance R_{PL}. The dynamic population $pop(t)$ was assumed to vary during the year with an unknown distribution with mean and variance estimates $\mathrm{E}[pop] = \frac{1}{T}\sum_{t=1}^{T} pop(t) = m_{pop}$ and $\mathrm{Var}[pop] = \mathrm{E}[pop^2] - \mathrm{E}[pop]^2 = R_{pop}$, respectively.

The population load, baseload, and person load were assumed to be statistically independent, which then gives the mean m_{tot} and variance R_{tot} of x_{tot} as

	$m_{tot} = \mathrm{E}[x_{tot}] = \mathrm{E}[x_{base}] + \mathrm{E}[pop]\mathrm{E}[x_{PL}] =$ $= m_{base} + m_{pop} m_{PL}$	(3a)
	$R_{tot} = \mathrm{Var}[x_{tot}] = \mathrm{Var}[x_{base}] + \mathrm{Var}[pop\ x_{PL}] =$ $= R_{base} + (\mathrm{Var}[pop] + \mathrm{E}[pop]^2)\,(\mathrm{Var}[x_{PL}] + \mathrm{E}[x_{PL}]^2) - (\mathrm{E}[pop]\mathrm{E}[x_{PL}])^2 =$ $= R_{base} + \left(R_{pop} + {m_{pop}}^2\right)\left(R_{PL} + {m_{PL}}^2\right) - \left(m_{pop} m_{PL}\right)^2,$	(3b)

where the law of expectations was used in (3a) and the variance of the product of independent variables in (3b).

We however assumed a correlation between the different person load components BOD, nitrogen, and phosphorus. That is, when the BOD person load is high, it is also likely that corresponding nitrogen and phosphorus person loads are high. Thus, the person loads (2c) were additionally assumed to follow a multivariate Gaussian as

	$X_{PL} \sim N(m_{PL}, R_{PL}),$	(4a)
	$M_{PL} = \begin{pmatrix} m_{x1_{PL}} \\ \cdots \\ m_{xN_{PL}} \end{pmatrix}$ $R_{PL} = \begin{pmatrix} R_{x1_{PL}} & .. & \sqrt{R_{x1_{PL}} R_{xN_{PL}}} \\ . & ... & .. \\ . & . & R_{xN_{PL}} \end{pmatrix}$	(b) (c)

with the covariances in the off-diagonal elements in (3c).

The load model (2) describe the underlying load drivers. In practice, several noise sources impact both the measurements and the underlying process. These have not been included here, but are straightforward to add, e.g., as additive Gaussian measurement noise.

4.3 Predicting a design load

The impact from population estimates on reactor sizing was exemplified by applying the four-step design process with the considered data.

1) Estimate the specific loads based on the static population estimate and mobility data.
2) Assume a predicted population of 330 000, which was provided from the municipal population prognosis.
3) Obtain the predicted population load as the product of 1) and 2)
4) Add additional industry load as the mean value 2019–2022 from the emission reports in Table 1.

A conventional WRRF with pre- and post-denitrification was then sized based on the predicted load using the automated and model-based approach in (Lindblom and Samuelsson, 2025). One thousand Monte Carlo design were conducted for each population data source to also include the variance in specific load.

5 Acknowledgements

This research was funded by Formas – a Swedish Research Council for Sustianable Development, and the strategic innovation program Smart Built Environment, grant ID U10-2022-13 and Digital Capacity 2025-00295. Project partners Uppsala Vatten och Avfall and Telia Company provided data and related expertise to the project. We gratefully acknowledge the staff that compiled the analysed data: Johanna Andersson, Jeanette Sipilä and David Lindhe from Uppsala Vatten och Avfall (wastewater load data and GIS data), Christian Lewenhaupt and Ludvig Uhmeier from Telia (mobility data and GIS data), and Christoffer Wärff from RISE and Beatrice Marselius from Uppsala Vatten och Avfall (daily load data).

6 References

Alex, J. (2024). Model-based construction of wastewater treatment plant influent data for simulation studies. Water 16(4), 564.

Baz-Lomba, J.A., Di Ruscio, F., Amador, A., Reid, M., Thomas, K.V., 2019. Assessing Alternative Population Size Proxies in a Wastewater Catchment Area Using Mobile Device Data. Environ. Sci. Technol. 53, 1994–2001.

Belia, E., Benedetti, L., Johnson, B., Murthy, S., Neumann, M., Vanrolleghem, P., Weijers, S. (Eds.), 2021. Uncertainty in Wastewater Treatment Design and Operation: Addressing current practices and future directions. IWA Publishing, London, UK.

Corominas, L., Zammit, I., Badia, S., Pueyo-Ros, J., Bosch, L.M., Calle, E., Martínez, D., Chesa, M.J., Chincolla, C., Martínez, A., Llopart-Mascaró, A., Varela, F.J., Domene, E., Garcia-Sierra, M., Garcia-Acosta, X., Satorras, M., Raich-Montiu, J., Peris, R., Horno, R., Rubión, E., Simón, S., Ribalta, M., Palacín, I., 2024. Profiling wastewater characteristics in intra-urban catchments using online monitoring stations. Water Science & Technology 89, 1512–1525.

Gernaey, K.V., Flores-Alsina, X., Rosen, C., Benedetti, L., Jeppsson, U., 2011. Dynamic influent pollutant disturbance scenario generation using a phenomenological modelling approach. Environmental Modelling & Software 26, 1255–1267.

Gudra, D., Dejus, S., Bartkevics, V., Roga, A., Kalnina, I., Strods, M., Rayan, A., Kokina, K., Zajakina, A., Dumpis, U., Ikkere, L.E., Arhipova, I., Berzins, G., Erglis, A., Binde, J., Ansonska, E., Berzins, A., Juhna, T., Fridmanis, D., 2022. Detection of SARS-CoV-2 RNA in wastewater and importance of population size assessment in smaller cities: An exploratory case study from two municipalities in Latvia. Science of The Total Environment 823, 153775.

Li, F., Vanrolleghem, P.A., 2022. An essential tool for WRRF modelling: a realistic and complete influent generator for flow rate and water quality based on data-driven methods. Water Science and Technology 85, 2722–2736.

Lindblom, E.U., Samuelsson, O., 2023. Comparison of guideline- and model-based WWTP design for uncertain influent conditions. Water Science and Technology 87, 218–227.

Lindblom, E.U., Samuelsson, O., 2025. Model-based sizing of wastewater resource recovery facilities with transparent safety factors. *[in preparation]*

Sim, W., Park, S., Ha, J., Kim, D., Oh, J.-E., 2023. Evaluation of population estimation methods for wastewater-based epidemiology in a metropolitan city. Science of The Total Environment 857, 159154.

Statistics Sweden, 2024. Folkmängden per månad efter kön, månad, region och ålder, Uppsala kommun (0380), 0-5 år 2019.

Statistics Sweden, 2020. Folkmängden för fastigheter inom Kungsängsverkets avrinningsområde 2020, dataexport från Uppsala Vatten och Avfall.

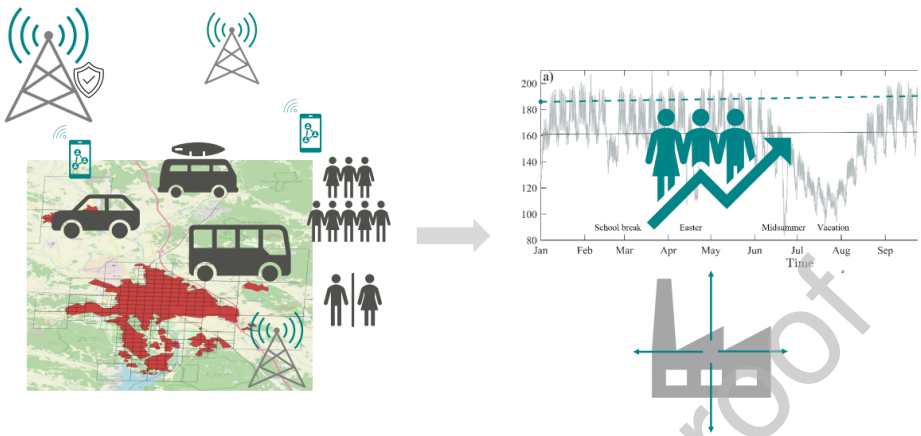
Talebizadeh, M., Belia, E., Vanrolleghem, P.A., 2016. Influent generator for probabilistic modeling of nutrient removal wastewater treatment plants. Environmental Modelling & Software 77, 32–49.

Uppsala universitet, 2023. Årsredovisning 2023.

Wärff, C., Carlsson, B., Arnell, M., Micolucci, F., Samuelsson, O., Jeppsson, U., 2025. Using a hybrid modelling approach for high time-resolution prediction of influent orthophosphate load in a water resource recovery facility. Water Research 286, 124176.

Yang, C., Belia, E., Daigger, G.T., 2022. Automating process design by coupling genetic algorithms with commercial simulators: a case study for hybrid MABR processes. Water Science and Technology 86, 672–689.

Graphical abstract



Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.