Wastewater catchment areas in Great Britain

Till Hoffmann*[1], Sarah Bunney[1], Barbara Kasprzyk-Hordern[2], and Andrew C. Singer[3]

[1]Department of Mathematics, Imperial College London

[2]Department of Chemistry, University of Bath

[3]UK Centre for Ecology and Hydrology

Wastewater catchment area data are essential for wastewater treatment capacity planning and have recently become critical for operationalising wastewater-based epidemiology (WBE) for COVID-19. Owing to the privatised nature of the water industry in the United Kingdom, the required catchment area datasets are not readily available to researchers. Here, we present a consolidated dataset of 7,537 catchment areas from ten sewerage service providers in the Great Britain, covering more than 96% of the population of England and Wales. We develop a geospatial method for estimating the population resident within each catchment from small area population estimates generated by the Office for National Statistics. The method is more widely applicable to matching electronic health records to wastewater infrastructure. Population estimates are highly predictive of population equivalent treatment loads reported under the European Urban Wastewater Treatment Directive. We highlight challenges associated with using geospatial data for wastewater-based epidemiology.

1 Introduction

Geospatial data on the extent of wastewater catchment areas are essential for wastewater-based epidemiology (WBE): information relating to the population resident in the catchment needs to be related to signals extracted from samples taken from the corresponding

*Corresponding author (t.hoffmann@imperial.ac.uk).

Code to reproduce the analysis is available at https://github.com/tillahoffmann/wastewater-catchment-areas.

wastewater treatment plant (WWTP) [6]. WBE has garnered recent attention because it can be used to monitor community transmission of SARS-CoV-2 relatively inexpensively as compared to mass testing [23], although with significant uncertainties [1, 29]. Evaluating the utility of wastewater-based monitoring of SARS-CoV-2 requires clinical data (such as case numbers, hospital admissions, or prevalence estimates [25]) aggregated to catchment areas for comparison. Beyond SARS-CoV-2, wastewater-based monitoring has the potential to provide further indicators of public health (such as dietary markers [10, 7], use of pharmaceuticals [27, 17, 30, 31], and other communicable diseases [2]). WBE has proven useful in estimating public exposure to food toxicants [26], lifestyle chemicals [4] and other hazardous chemicals [18, 14]. Diverse methods for estimating the population size served by individual WWTPs have been developed, including the use of biochemical markers [5] and mobile phone data [3]. But population estimation remains a challenging problem. Therefore, irrespective of the particular application, catchment area data are required to relate the signal to other data (such as prescribing data or the census [28]).

Due to the privatisation of the water sector in England in 1989, catchment area data are held by different companies and are not available in a singular coherent form. This lack of data hampers research efforts and poses challenges for open and reproducible research. To fill this data gap, we obtained catchment area data from sewerage service providers in the United Kingdom under the Environmental Information Regulations 2004 which provide a statutory right to access environmental information [12]. We consolidated the data, removed duplicate catchments, annotated conflicting information, and matched wastewater catchments to the respective treatment plant and associated metadata. The metadata include population equivalent treatment load (measured by biological oxygen demand), as reported under the Urban Wastewater Treatment Directive of the Council of the European Union [9]. We demonstrate the utility of wastewater catchment area data by estimating the population resident within each catchment based on small-area population estimates produced by the Office for National Statistics (ONS) [20], the official body generating and publishing population statistics in the United Kingdom. These estimates are highly predictive of the population equivalent load of treatment works, indicating that geospatial approaches are a useful tool for connecting wastewater-based signals with external datasets.

2 Methods

2.1 Data and data preparation

2.1.1 Wastewater catchment area data

We submitted Environmental Information Requests to the twelve major sewerage service providers in the United Kingdom, and ten providers supplied catchment area data, as shown in table 1. An overview of the catchments is shown in fig. 1 (a). Thames Water, United Utilities, and Wessex Water respectively provided 4,167, 1,292, and 402 subcatchments, i.e. a breakdown of the catchment area serviced by a single treatment

Company	Catchments		Area (km^2)	Matched UWWTPs	Population estimate
	Provided	Retained			
Anglian	1,149	1,140	4,207	317	6.9m
Northern Ireland	data not provided				
Northumbrian	322	320	1,304	66	2.7m
Scottish	1,877	1877	1,974	198	—
Severn Trent	1,016	1,014	2,485	260	9.3m
Southern	383	379	1,402	139	4.8m
Southwest	data not provided				
Thames	353	349	2,652	156	14.6m
United Utilities	569	567	1,935	147	7.2m
Welsh	866	864	1,172	126	3.4m
Wessex	385	385	2,139	117	2.9m
Yorkshire	617	617	1,693	151	5.1m
Total	7,537	7,512	20,962	1,677	56.8m

Table 1: Overview of data provided by different sewerage service providers in the United Kingdom under Environmental Information Requests and 2016 population estimates (see section 2.2 for details). The number of Urban Wastewater Treatment Plants (UWWTPs) is smaller than the number of catchments because data are only reported for treatment plants with a BOD population equivalent of more than 2,000. Areas covered and geospatial population estimates may not add up to the total due to rounding.

plant into smaller areas. Data were aggregated by considering the spatial union of all subcatchments that drain to the same treatment plant. Severn Trent Water provided catchment area data for 1,016 of their own treatment plants as well as 58 catchments of partner companies; we removed the latter from the dataset to avoid duplication. Southwest Water did not provide catchment area data because they do not consider the geospatial extent of wastewater catchment areas environmental information; Northern Ireland Water did not provide catchment area data because "the information requested is currently considered sensitive".

To assess the integrity of the dataset, we manually reviewed 87 pairs of intersecting catchments from different providers. We removed 25 catchments that were supplied in duplicate, as shown in fig. 1 (b). For some pairs, it was not possible to uniquely determine which of the overlapping catchments services a given area, as shown in fig. 1 (c). We thus annotated both catchments for future reference, resulting in 52 catchments with annotations[1].

2.1.2 Urban Wastewater Treatment Directive regulatory data

Under article 17 of the Urban Wastewater Treatment Directive (UWWTD) adopted in 1991, members states are obliged to report details on urban wastewater treatment plants every two years [8]. Reporting standards were formalised in 2014 [9], and data need to be provided for treatment plants exceeding a population equivalent of 2,000. Article 2 defines a population equivalent (p.e.) as "organic biodegradable load having a five-day biochemical oxygen demand (BOD5) of 60 g of oxygen per day" [8]. We henceforth refer to population equivalents reported under the UWWTD as BOD p.e. These data include both the actual load entering and the treatment capacity of each treatment plant (both measured as BOD p.e.) from 2006 to 2016 together with the GPS coordinates of the treatment plant [13]. The number of records per year varies between 1,852 and 1,908 over the reporting period; the dataset includes a total of 1,990 unique treatment plants.

To match the UWWTD treatment plants to the catchment areas described in section 2.1.1, we removed 133 treatment works that were labelled as "inactive" in their latest report because catchment data were obtained in 2021, leaving 1,857 unique treatment plants. We also ignored 170 treatment plants located in the service area of Southwest Water, Northern Ireland, and Gibraltar because corresponding catchment area data are not available, leaving 1,687 unique treatment plants. We evaluated the distance between all pairs of catchment areas and the most recently reported location of treatment plants. We automatically matched 1,492 treatment plants to catchments if the distance between them was less than 100 metres and there were no other treatment plants within 100 metres of the same catchment. We also matched 95 treatment plants to catchments that were less than 2,500 metres apart and whose names matched exactly after removing special characters, whitespace, and common acronyms (such as "WWTW" or "STP"). In total, 1,587 treatment plants were matched automatically, leaving 100 unmatched treatment plants which we reviewed manually to match them to catchments. After manual

[1]The number of removed and annotated catchments do not add up to the number of pairs because a single catchment may overlap with multiple segments from other providers.

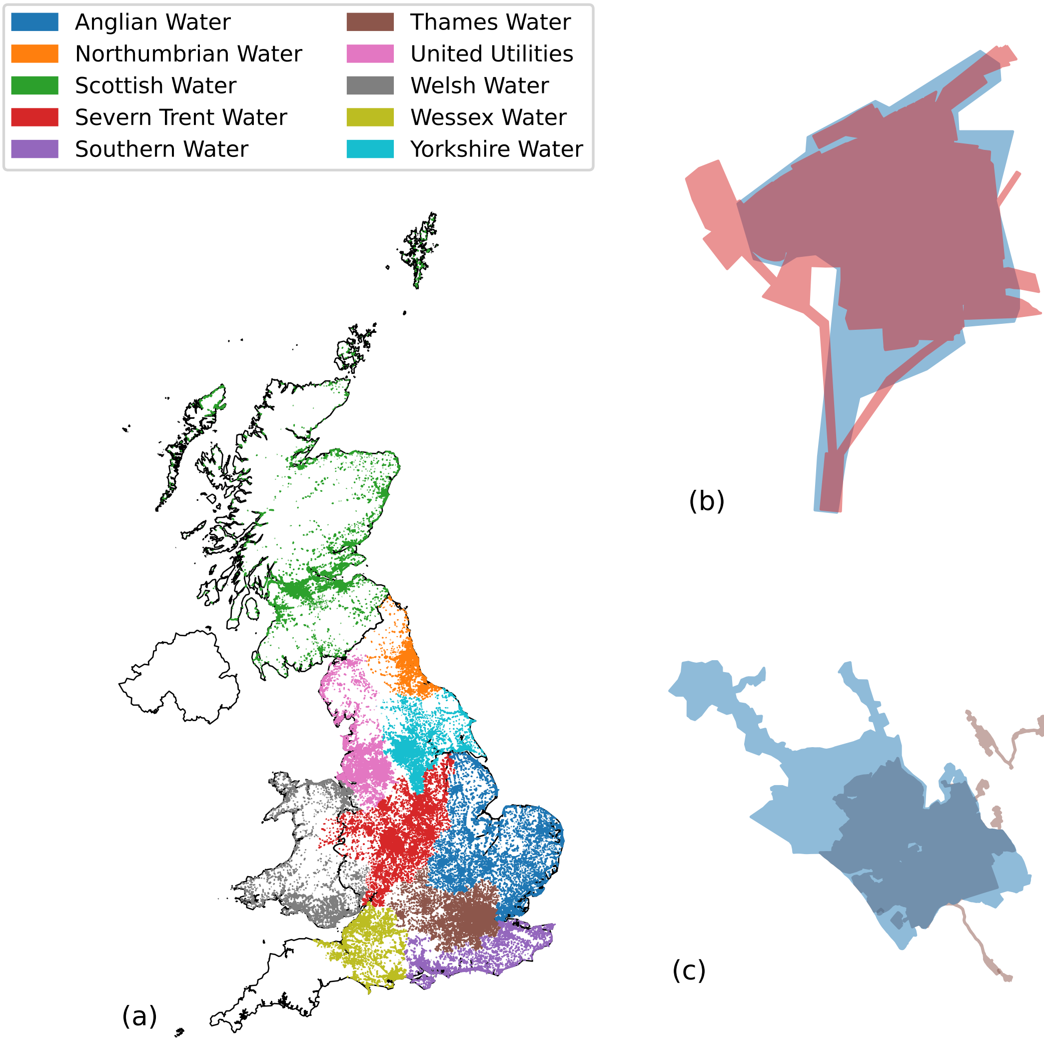

Figure 1: A consolidated dataset covers 7,537 catchment areas from ten sewerage service providers, as shown in panel (a). Catchment area data supplied by different providers may conflict when different catchments cover the same geographical area. Panel (b) shows an example of Market Overton catchments being supplied in duplicate by different providers. Market Overton is serviced by Severn Trent Water, and the version provided by Anglian Water has been removed from the consolidated dataset. Panel (c) shows an example of two conflicting catchments, East Hyde (Thames Water) and Chalton (Anglian Water), that cannot easily be resolved, resulting in annotations in the consolidated dataset.

review, 11 treatment plants could not be matched to a catchment. The median distance between matched treatment plants and catchments is zero, and the 99$^{\text{th}}$ percentile is 537 metres.

2.1.3 Census boundaries and small area population estimates

We obtained geospatial boundaries of lower-layer super output areas (LSOAs) for the 2011 census from the ONS [19]. These 34,753 areas cover England and Wales and are used as statistical reporting units. For each LSOA, we obtained population estimates for the period between 2001 and 2017 [20]. These small-area population estimates (SAPEs) are mid-year estimates of the typical resident population and do not account for transient populations, such as commuters or holiday makers [21]. Census data are used to generate SAPEs in census years, and, for intercensal years, they are generated by "rolling forward" the resident population based on changes in the number of people registered with general practitioners in the area [21]. See Park [21] for detailed methodology for SAPEs and Park [22] for a discussion on the quality of SAPEs.

2.2 Geospatial population estimates

We considered the spatial overlap between catchment areas and LSOAs to generate geospatial population estimates (GPEs) for each catchment area. In particular, let c_i be the extent of catchment i and l_j be the extent of LSOA j; we use $|x|$ to denote the area of an entity x. The overlap of a catchment i and LSOA j is

$$a_{ij} = c_i \cap l_j,$$

where we have used set notation to denote the intersection. For each pair, we assign a proportion $|a_{ij}| / |A_j|$ of the population n_j of LSOA j to catchment i, where

$$A_j = \bigcup_i a_{ij}$$

is the subset of LSOA j serviced by *any* catchment area. In contrast to Tscharke, O'Brien, Ort, Grant, Gerber, Bade, Thai, Thomas, and Mueller [28], we used the area serviced in the denominator (as opposed to the total area of the LSOA) because LSOAs are census reporting areas that cover the United Kingdom and may extend into unserviced areas, such as green spaces. For example, fig. 2 shows LSOA E01003817 in south west London that partially covers Richmond Park. The entire population is resident near the boundary of the LSOA, and the majority of the area of the LSOA is uninhabited. Using the total area of the LSOA in the denominator would underestimate the population contribution to both the Mogden and Hogsmill treatment works in west London, servicing approximately two million and 400 thousand people, respectively. The population estimate m_i for catchment area i is thus

$$m_i = \sum_j \frac{|a_{ij}|}{|A_j|} n_j.$$

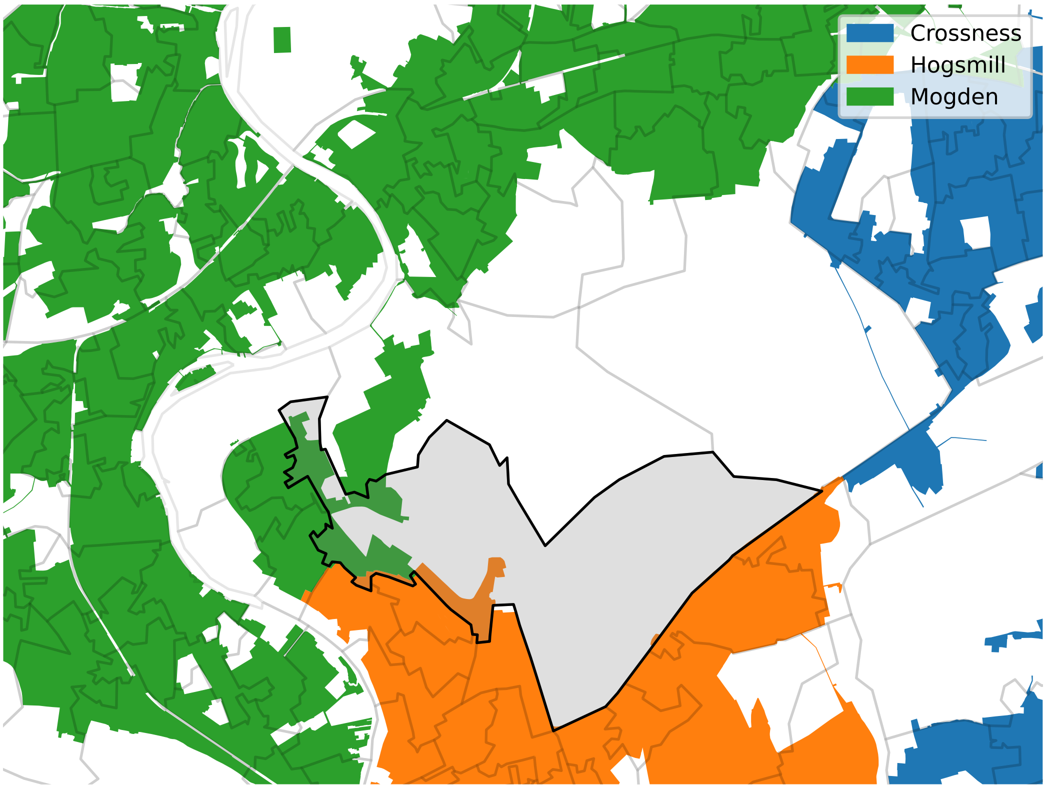

Figure 2: *Different normalisation options can yield substantially different population estimates.* Three different catchment areas in south west London are shown as solid shapes; the large central white region is Richmond Park which is not serviced by any catchment area. LSOA boundaries are shown as grey lines, and LSOA E01003817 is shown as a shaded region with black boundary. Assigning population estimates to a catchment based on the proportion of the *total area* that overlaps with the catchment would underestimate the serviced population. We instead normalise by the *area covered by any catchment* which guarantees that every person is accounted for in population estimates.

This method can also be used to connect electronic health records, such as COVID-19 case data or prescribing data, with wastewater infrastructure data.

3 Results

Geospatial population estimates are highly predictive of population equivalent loads reported under the UWWTD, as shown in fig. 3 (a). The Pearson correlation on the log-scale is 0.977 with a p-value $< 10^{-6}$. Surprisingly, geospatial population estimates and BOD p.e. are not only correlated, but their dependence closely follows the identity relation despite geospatial population estimates only accounting for domestic contributions. Unless substantial wastewater volumes are treated in private treatment plants, this suggests that treatment plants in the United Kingdom have little residual capacity, consistent with repeated discharge of untreated wastewater into the natural environment during adverse weather events [15].

Outliers highlight limitations of geospatial approaches to population estimation that would also affect wastewater-based epidemiology. For example, the Haggerston sewage treatment plant in Northumberland has a BOD p.e. load of over 2,000, but the GPE is almost an order of magnitude smaller because the treatment plant services a holiday park with large transient population that is not captured by the census. In contrast, the geospatial population estimate for the Billericay treatment plant in Essex is larger than the reported BOD p.e. load. This is a result of part of the wastewater being redirected to the nearby Shenfield and Hutton treatment plant to alleviate pressure on the Billericay treatment plant caused by recent housing developments [24]. The Rotherwas treatment plant in Hereford has a BOD p.e. load more than 30 times larger than the estimated population, in part due to industrial influent from the Bulmers cider factory and pumping from the nearby Eign treatment plant [16]. But discrepancies between geospatial population estimates and BOD p.e. load can also be the consequence of data quality issues, as illustrated by the Chalton treatment plant: as shown in fig. 1 (c), the Chalton catchment (Anglian Water) overlaps substantially with the East Hyde catchment (Thames Water), and it is not possible to determine which treatment work services the population.

To summarise the relationship between BOD p.e. load and geospatial population estimates, we calculated the median absolute error (MAE) across treatment plants on the $\log_{10}$ scale, i.e.

$$\mathrm{MAE} = \mathrm{median}_i \left| \log_{10} \left(\frac{m_i}{p_i} \right) \right|,$$

where p_i is the BOD p.e. load reported under the UWWTD. As shown in fig. 3 (b), the MAE decreases over time. This is expected as the wastewater catchment area data were obtained in 2021 and may thus not reflect wastewater infrastructure from more than a decade ago well. The reduction may also be the result of improvements in reporting which was formalised in 2014 [9]. Irrespectively, the MAE between 0.046 and 0.051 is small across the entire period from 2006 to 2016 (corresponding to relative errors of

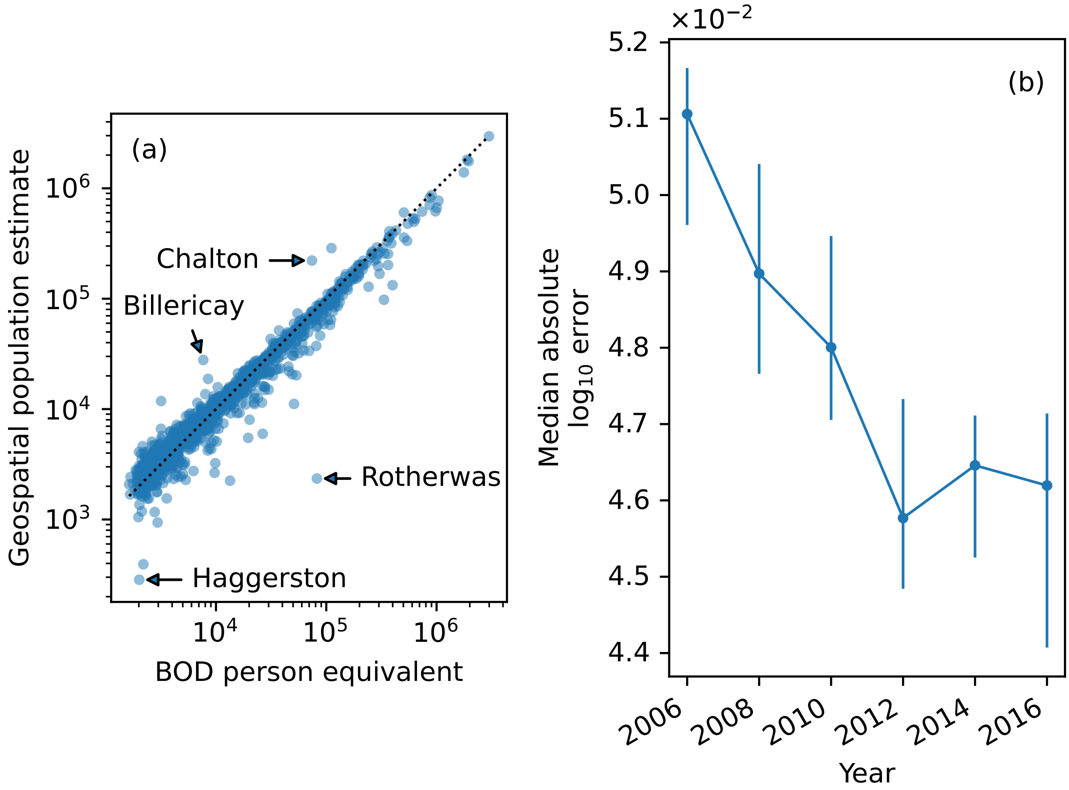

Figure 3: *Geospatial population estimates (GPEs) are highly predictive of BOD population equivalent (p.e.) treatment load.* Panel (a) shows a scatter of GPEs based on 2016 mid-year population estimates against treatment load from UWWTD data in 2016. Four outliers are highlighted with additional context in the main text. Panel (b) shows the median absolute $\log_{10}$ error between geospatial population estimates and BOD p.e. load over time. Error bars correspond to the bootstrapped interquartile range.

approximately 11%) despite geospatial population estimates and BOD p.e. load being conceptually different quantities.

4 Discussion

We have presented a consolidated dataset of 7,537 wastewater catchment areas with wide applicability for wastewater-based epidemiology and water research in Great Britain. To demonstrate the utility of the dataset, we used small-area population estimates from the Office for National Statistics to estimate the number of people resident within each catchment and compared the results with population equivalent treatment loads reported under the Urban Wastewater Treatment Directive. Geospatial population estimates (GPEs) are highly predictive but also highlight limitations of geospatial approaches: first, GPEs based on census data cannot capture transient populations, such as holiday parks as illustrated by the treatment plant in Haggerston. Second, GPEs are affected by uncertainties pertaining to small area population estimates (especially for intercensal years) as well as the mapping of census reporting areas (such as LSOAs) to catchment areas. Third, understanding pumping between treatment plants to alternative sites (such as from Billericay to Shenfield and Hutton) requires additional data, such as network information and pumping stations which are not currently available. Fourth, private sewage treatment works and septic tanks, which are more common in rural areas, are not accurately captured by this dataset: wastewater may be tankered to different treatment works or treated on site. Data on private sewage treatment works are not readily available because most of them do not need to be registered with the Environment Agency [11]. Finally, agricultural and industrial discharges as well as tankered waste can give rise to large treatment loads that are not reflected in population estimates (such as influent from the Bulmers factory at Rotherwas). In short, wastewater treatment infrastructure is a complex system whose idiosyncracies cannot be captured solely by geospatial data. Despite these challenges, we believe the data and methods presented here can support water research in general and wastewater-based epidemiology in particular.

Acknowledgements

We thank Anglian Water, Northumbrian Water, Scottish Water, Severn Trent Water, Southern Water, United Utilities, Welsh Water, Wessex Water, and Yorkshire Water for providing catchment area data in response to requests under the Environmental Information Regulations 2004. This research is part of the Data and Connectivity National Core Study, led by Health Data Research UK in partnership with the Office for National Statistics and funded by UK Research and Innovation (grant ref MC_PC_20029).

References

[1] W. Ahmed et al. “Minimizing Errors in RT-PCR Detection and Quantification of SARS-CoV-2 RNA for Wastewater Surveillance”. In: *Sci. Total Environ.* 805 (2022), p. 149877. DOI: 10.1016/j.scitotenv.2021.149877.

[2] H. Asghar, O. M. Diop, G. Weldegebriel, F. Malik, S. Shetty, L. El Bassioni, A. O. Akande, E. Al Maamoun, S. Zaidi, A. J. Adeniji, C. C. Burns, J. Deshpande, M. S. Oberste, and S. A. Lowther. “Environmental Surveillance for Polioviruses in the Global Polio Eradication Initiative”. In: *J. Infect. Dis.* 210 (Nov. 2014), S294–S303. DOI: 10.1093/infdis/jiu384.

[3] J. A. Baz-Lomba, F. Di Ruscio, A. Amador, M. Reid, and K. V. Thomas. “Assessing Alternative Population Size Proxies in a Wastewater Catchment Area Using Mobile Device Data”. In: *Environ. Sci. Technol.* 53.4 (2019), pp. 1994–2001. DOI: 10.1021/acs.est.8b05389.

[4] J. A. Baz-Lomba et al. “Comparison of pharmaceutical, illicit drug, alcohol, nicotine and caffeine levels in wastewater with sale, seizure and consumption data for 8 European cities”. In: *BMC Public Health* 16.1 (2016), p. 1035. DOI: 10.1186/s12889-016-3686-5.

[5] F. Been, L. Rossi, C. Ort, S. Rudaz, O. Delémont, and P. Esseiva. “Population Normalization with Ammonium in Wastewater-Based Epidemiology: application to Illicit Drug Monitoring”. In: *Environ. Sci. Technol.* 48.14 (2014), pp. 8162–8169. DOI: 10.1021/es5008388.

[6] P. M. Choi, B. J. Tscharke, E. Donner, J. W. O’Brien, S. C. Grant, S. L. Kaserzon, R. Mackie, E. O’Malley, N. D. Crosbie, K. V. Thomas, and J. F. Mueller. “Wastewater-based epidemiology biomarkers: past, present and future”. In: *Trends Anal. Chem.* 105 (2018), pp. 453–469. DOI: 10.1016/j.trac.2018.06.004.

[7] P. M. Choi, B. Tscharke, S. Samanipour, W. D. Hall, C. E. Gartner, J. F. Mueller, K. V. Thomas, and J. W. O’Brien. “Social, demographic, and economic correlates of food and chemical consumption measured by wastewater-based epidemiology”. In: *Proc. Natl. Acad. Sci. U.S.A* 116.43 (2019), pp. 21864–21873.

[8] Council of the European Union. “91/271/EEC”. In: *Offic. J. Eur. Communities* 135 (1991), pp. 40–52.

[9] Council of the European Union. “91/271/EEC”. In: *Offic. J. Eur. Communities* 197 (2014), pp. 77–86.

[10] A. J. Cross, J. M. Major, and R. Sinha. “Urinary Biomarkers of Meat Consumption”. In: *Cancer Epidemiol. Biomark. Prev.* 20.6 (2011), pp. 1107–1111. DOI: 10.1158/1055-9965.EPI-11-0048.

[11] Department for Environment, Food, and Rural Affairs and Environment Agency. *Reform of the regulatory system to control small sewage discharges from septic tanks and small sewage treatment plants in England*. 2014. URL: https://www.gov.uk/government/publications/small-sewage-discharges-in-england-general-binding-rules.

[12] *Environmental Information Regulations*. 2004. URL: https://www.legislation.gov.uk/uksi/2004/3391/contents/made.

[13] European Environment Agency. *Waterbase: Urban Wastewater Treatment Directive Reported Data*. 2020.

[14] I. González-Mariño et al. "Assessing population exposure to phthalate plasticizers in thirteen Spanish cities through the analysis of wastewater". In: *J. Hazard. Mater.* 401 (2021), p. 123272. DOI: 10.1016/j.jhazmat.2020.123272.

[15] P. Hammond, M. Suttie, V. T. Lewis, A. P. Smith, and A. C. Singer. "Detection of untreated sewage discharges to watercourses using machine learning". In: *npj Clean Water* 4.1 (2021). DOI: 10.1038/s41545-021-00108-3.

[16] Hereford Times. *Welsh Water to spend £1.2m at sewage works*. 2001. URL: https://www.herefordtimes.com/news/5704789.welsh-water-to-spend-12m-at-sewage-works/.

[17] B. Kasprzyk-Hordern, K. Proctor, K. Jagadeesan, L. Lopardo, K. J. O'Daly, R. Standerwick, and R. Barden. "Estimation of community-wide multi-chemical exposure via water-based chemical mining: key research gaps drawn from a comprehensive multi-biomarker multi-city dataset". In: *Environ. Int.* 147 (2021), p. 106331. DOI: 10.1016/j.envint.2020.106331.

[18] L. Lopardo, B. Petrie, K. Proctor, J. Youdan, R. Barden, and B. Kasprzyk-Hordern. "Estimation of community-wide exposure to bisphenol A via water fingerprinting". In: *Environ. Int.* 125 (2019), pp. 1–8. DOI: 10.1016/j.envint.2018.12.048.

[19] Office for National Statistics. *Lower Layer Super Output Areas (December 2011) Boundaries, Generalised Clipped*. 2021.

[20] Office for National Statistics. *Population estimates for lower layer super output areas (LSOA) in England and Wales, single year of age and sex, mid-2001 to mid-2017*. 2019.

[21] N. Park. *Methodology note on production of population estimates by output areas, electoral, health and other geographies, England and Wales*. Tech. rep. Office for National Statistics, 2021.

[22] N. Park. *Population estimates by output areas, electoral, health and other geographies QMI*. Tech. rep. Office for National Statistics, 2021.

[23] D. Polo, M. Quintela-Baluja, A. Corbishley, D. L. Jones, A. C. Singer, D. W. Graham, and J. L. Romalde. "Making waves: wastewater-based epidemiology for COVID-19 – approaches and challenges for surveillance and prediction". In: *Water Res.* 186 (2020), p. 116404. DOI: 10.1016/j.watres.2020.116404.

[24] C. Postlethwaite, C. Pelling, R. Sweet, and J. Robinson. *South Essex Water Cycle Study.* Tech. rep. URS/Scott Wilson, 2011.

[25] K. B. Pouwels et al. "Community prevalence of SARS-CoV-2 in England from April to November, 2020: results from the ONS Coronavirus Infection Survey". In: *Lancet Public Health* 6.1 (2021), e30–e38. DOI: 10.1016/S2468-2667(20)30282-6.

[26] N. I. Rousis et al. "Wastewater-based epidemiology to assess pan-European pesticide exposure". In: *Water Res.* 121 (2017), pp. 270–279. DOI: 10.1016/j.watres.2017.05.044.

[27] L.-H. Sheng, H.-R. Chen, Y.-B. Huo, J. Wang, Y. Zhang, M. Yang, and H.-X. Zhang. "Simultaneous Determination of 24 Antidepressant Drugs and Their Metabolites in Wastewater by Ultra-High Performance Liquid Chromatography–Tandem Mass Spectrometry". In: *Molecules* 19.1 (2014), pp. 1212–1222. DOI: 10.3390/molecules19011212.

[28] B. J. Tscharke, J. W. O'Brien, C. Ort, S. Grant, C. Gerber, R. Bade, P. K. Thai, K. V. Thomas, and J. F. Mueller. "Harnessing the Power of the Census: characterizing Wastewater Treatment Plant Catchment Populations for Wastewater-Based Epidemiology". In: *Environ. Sci. Technol.* 53.17 (2019), pp. 10303–10311. DOI: 10.1021/acs.est.9b03447.

[29] M. J. Wade et al. "Understanding and managing uncertainty and variability for wastewater monitoring beyond the pandemic: lessons learned from the United Kingdom national COVID-19 surveillance programmes". In: *J. Hazard. Mater.* 424 (2022), p. 127456. DOI: 10.1016/j.jhazmat.2021.127456.

[30] Y. Xiao, X.-T. Shao, D.-Q. Tan, J.-H. Yan, W. Pei, Z. Wang, M. Yang, and D.-G. Wang. "Assessing the trend of diabetes mellitus by analyzing metformin as a biomarker in wastewater". In: *Sci. Total Environ.* 688 (2019), pp. 281–287. DOI: 10.1016/j.scitotenv.2019.06.117.

[31] Y. Zhang, L. Duan, B. Wang, Y. Du, G. Cagnetta, J. Huang, L. Blaney, and G. Yu. "Wastewater-based epidemiology in Beijing, China: prevalence of antibiotic use in flu season and association of pharmaceuticals and personal care products with socioeconomic characteristics". In: *Environ. Int.* 125 (2019), pp. 152–160. DOI: 10.1016/j.envint.2019.01.061.