

Population Socioeconomics Predicted Using Wastewater

Phil M. Choi, Jake W. O'Brien, Ben J. Tschärke, Jochen F. Mueller, Kevin V. Thomas, and Saer Samanipour*



Cite This: *Environ. Sci. Technol. Lett.* 2020, 7, 567–572



Read Online

ACCESS |



Metrics & More

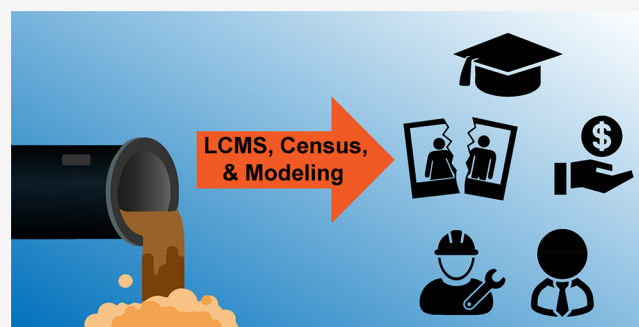


Article Recommendations



Supporting Information

ABSTRACT: Municipal wastewater typically contains many drugs and anthropogenic chemicals or biomarkers. The occurrence of these chemicals in wastewater is linked to the socioeconomic characteristics of the contributing population. Based on these relationships, we propose, execute and evaluate a novel model for predicting population socioeconomics. Specifically, we used biomarkers in wastewater to predict 37 socioeconomic characteristics of populations during the Australian Census. The resultant model was further tested on nine other populations separate from the training data set. Prediction performance in the test populations (defined as accuracy \pm SD) fit within 75% and 125% for many features such as catchment median age, and specific measures of educational attainment (e.g., high school completion) and employment (e.g., managerial employment). Considering the relative ease, low cost and high frequency at which wastewater samples can be collected and analyzed, wastewater analysis could be used as a complementary technique for assessing population socioeconomics.



INTRODUCTION

Many states use population censuses to understand key characteristics of their populations. Census results are greatly valued by both government and nongovernment bodies in planning and evaluating policies, services, and infrastructure. However, censuses are resource intensive. The 2016 Australian Census cost \$514 million USD (\$22 per capita),¹ while the 2010 US Census cost \$13 billion USD (\$42 per capita).² They are typically conducted infrequently, such as once every five years in Australia and once every 10 years in the USA, UK, China, and India. Some countries have only recently begun regular censuses (e.g., Germany), and many do not hold a regular census.

Chemical consumption patterns within a population can reflect various population characteristics.³ Wastewater-based epidemiology (WBE) is capable of measuring such chemical consumption patterns at a community or population scale. Chemicals can be measured from wastewater systematically sampled (24 h composites) from the inlet of a wastewater treatment plant (WWTP).^{4,5} Measurements can be expressed in terms of mass of chemical excreted per capita per day by normalizing to wastewater flow and catchment population size.^{6–8}

WBE studies performed in tandem with a census have highlighted strong links between population normalized chemical loads and socioeconomic characteristics of the wastewater catchments. A recent WBE publication showed significant correlations between diet biomarkers and pharmaceuticals and personal care products (PPCPs) with various

aspects of socioeconomics, such as median age, socioeconomic index, and educational attainment.³ WWTP population size alone had strong correlations with per capita loads of various PPCPs, especially acesulfame ($R^2 = 0.995$) and gabapentin ($R^2 = 0.968$).⁹

While there are many established links between wastewater biomarkers and socioeconomic characteristics, these relationships have only been featured in simple observational studies.^{3,10,11} Here, we assessed whether WBE biomarkers could be used to predict various socioeconomic features of populations (Table S1). Per capita mass loads of 40 biomarkers and 37 socioeconomic features were used to develop a partial least-squares model. The model was subsequently evaluated by testing on other populations. This study presents a novel predictive modeling approach to evaluate the potential socioeconomic characteristics of populations via WBE. This approach could be adopted as a novel method for surveying socioeconomic changes in populations in a manner that is much cheaper and faster than traditional large-scale survey based methods.

Received: May 13, 2020

Revised: June 26, 2020

Accepted: June 26, 2020

Published: June 26, 2020



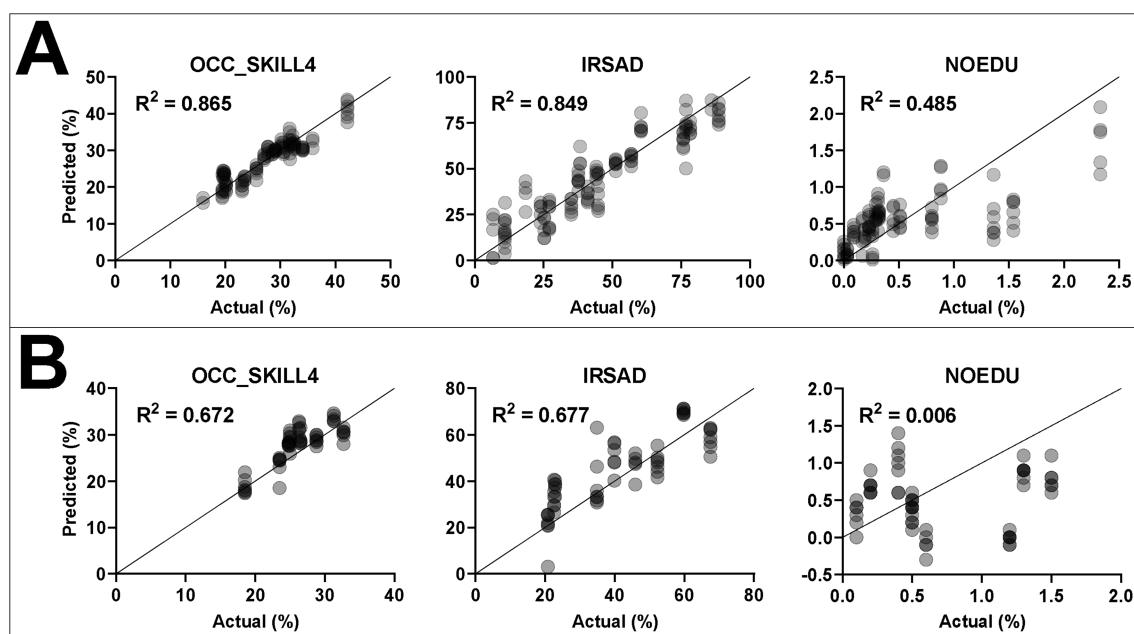


Figure 1. Agreement between actual and predicted socioeconomic features in the training data set (A) and test data set (B). Three representative features are shown: OCC_SKILL4 (individuals employed in jobs commensurate with vocational training or one year of relevant experience), IRSAD (index of relative socioeconomic advantage or disadvantage), and NOEDU (individuals without formal education). Lines shown are $x = y$.

MATERIALS AND METHODS

Wastewater Samples. All wastewater samples in this study were 24 h time- or flow-composited influent from WWTPs (Table S3).¹² For each WWTP, samples were collected during five to seven consecutive days during the week of the 2016 Australian Census, such that one wastewater sample from each WWTP corresponded to census night (August 9, 2016). Each archived wastewater sample was filtered with regenerated cellulose filters (0.2 μm). A 1 mL aliquot was spiked with internal standard mix prior to analysis. Biomarkers (Table S2) were measured from wastewater samples using standard liquid chromatography tandem mass spectrometry processes detailed and validated elsewhere.^{3,13}

Census Data. Each WWTP catchment was matched with population size and 37 socioeconomic features (including median population age and socioeconomic index, Table S1) for the population residing in each WWTP catchment during census night. Detailed methods are available elsewhere.^{3,7} Briefly, census results were downloaded using TableBuilder Pro (Australia Bureau of Statistics), and socioeconomic features were calculated according to their respective socioeconomic indexes for areas (SEIFA) formulas.¹⁴ Spatial resolution was at mesh block (avg. 65 capita/mesh block) level for population size and median age and index of relative socioeconomic advantage or disadvantage (IRSAD) and at SA1 (avg. 400 capita/SA1) level for all other features. The socioeconomic features tailored to each WWTP catchment were calculated using georeferenced maps of WWTPs in R Studio (v3.5.2).

Data Sets. We used a training data set of 40 WBE biomarkers (six licit drugs, two illegal drugs, seven opioids, eight antidepressants and psychotics, nine other pharmaceuticals, two artificial sweeteners, and six diet biomarkers) from an existing publication³ for model training and validation. This data set was acquired in 2018. In addition, a separate test data set was acquired for separate WWTP catchments to evaluate

the model for catchments not included in the training set (Figure S2). This test data set was acquired using identical analytical methods used for the training data set but was analyzed 9 months apart. The training set included the X block (independent variables) composed of per capita loads of 40 biomarkers in 142 wastewater samples (from 22 WWTPs in six states and territories) and a Y block (dependent variables) populated with 37 socioeconomic features for each WWTP (Table S4). The test set, on the other hand, had the same structure and included the same variables measured in 57 wastewater samples (from nine WWTPs in five states and territories) (Table S5). There was no overlap in WWTP catchments of the training and test sets. This enabled us to assess the general applicability of our model across Australia. The test and training data sets represented 23.9% of the Australian population.

Statistical Analysis. We developed a PLS regression¹⁵ in Matlab (R2015) to model and predict different socioeconomic parameters based on per capita loads of biomarkers. The training set was used for model building and cross validation. All data sets were first mean-centered and Pareto scaled.¹⁶ The model was initialized using 20 components. We calculated the cumulative variance explained by the model and root-mean-square error (RMSE) as a function of the number of components as well as the regression coefficient (i.e., R^2) of the model. The model with three components was selected to avoid overfitting and went through 10 folds of cross-validation. This process consisted of two independent steps first dividing the data into 10 randomly selected groups and using the nine data sets as the training set, while the tenth one is used as the validation set. This process was repeated until all the 10 subdata sets are used as the validation set at least once. The second step was to generate randomly distributed X and Y blocks and perform the regression between these simulated blocks for a total of 1000 simulations. Furthermore, we randomly sampled 75% of the training set and used these data to regenerate the three-component model. We repeated this

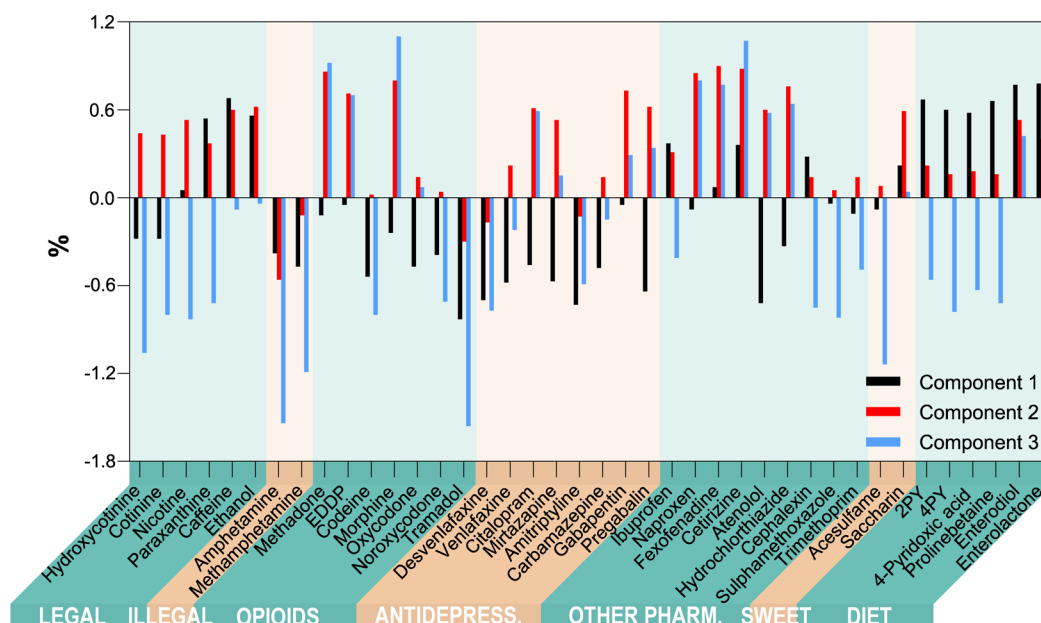


Figure 2. Weightings of PLS components 1, 2, and 3, which described 48%, 9%, and 8% of the variance in the data, respectively. Biomarkers have been grouped into legal drugs, illegal drugs, opioids, antidepressants and psychotics, other pharmaceuticals, artificial sweeteners, and diet biomarkers. EDDP, 2-ethylidene-1,5-dimethyl-3,3-diphenylpyrrolidine; 2PY, N1-methyl-2-pyridone-5-carboxamide; 4PY, N1-methyl-4-pyridone-3-carboxamide.

step 500 times and used the coefficients of these 500 models for defining the uncertainty of the model as well as to evaluate its robustness. Finally, we used the test set to assess the model applicability for catchments outside the training set.

RESULTS AND DISCUSSION

Model Characteristics. Three components in our PLS model explained 63% of the variance in the training data set (Figure S1). This validated model produced a median (i.e., over 500 simulations) regression coefficient (R^2) of 0.80 over the 37 socioeconomic features, having a statistical significance of greater than 98% (calculated based on the R^2 of random distributions during the cross-validation) (Figure 1A).^{15,17} Median RMSE was 2.55% (range: 0.52%–8.51%), indicating relatively small discrepancies between predicted and observed values. While many features such as OCC_SKILL4 (individuals employed in jobs commensurate with vocational training or one year of relevant experience) and IRSAD were well predicted by the model, several features such as NOEDU (no formal education) were less successfully predicted. This was ascribed to a high proportion of zero or near-zero entries in the data sets (Figure 1, Table S6). Zero entries reduce variable space and predictive power. Near-zero entries added uncertainty since small numerical values in census results are randomly adjusted to protect privacy and are additionally prone to respondent and processing errors.¹⁸

The first component of the model described 48% of the variance in the data and was characterized by negative weightings for caffeine and alcohol biomarkers, opioids and antidepressants, and positive weightings for vitamin and citrus dietary fiber consumption biomarkers (Figure 2). The magnitudes of weightings for the first component bear a strong resemblance to the correlations of each biomarker (or biomarker group) with IRSAD,³ which further supports model validity. The second component, covering 9% of variance in the data, had positive weightings for legal recreational drugs,

antipsychotics, antihistamines, and antihypertensives. The third component covered 8% of the variance and had mixed influences from licit and illicit drugs, certain opioids, antihistamines, and acesulfame. Overall, these results indicate that a large portion of socioeconomic was explained by diet, caffeine, alcohol consumption, and other behaviors linked to the consumption of pharmaceuticals used to treat various aspects of “distress”.

Model Assessment Using Test Data Set. The model was used to predict socioeconomic features for catchments outside of the training set (i.e., the test set, Figure S2). Prediction performance for each variable is reported by dividing the prediction error by the actual census determined value. Each variable was categorized as good (prediction \pm SD within 75% and 125% of census value), adequate (prediction \pm SD within 0% and 200%), or poor (prediction \pm SD $<$ 0% and $>$ 200%).

As shown in Figure 3, prediction was good or adequate for features relating to education such as DIPLOMA (highest qualification is an advanced diploma), DEGREE (bachelor degree or higher), and NOYEAR12ORHIGHER (no year/grade 12 education or higher). This may be because educational attainment is closely related to diet quality^{19,20} and risk taking behavior,²¹ aspects of which were captured by the biomarkers of diet (vitamin consumption, fiber consumption, and citrus) and illegal drug use (amphetamine, methamphetamine) in this study (Table S2). Prediction was also good or adequate for all but one of the occupational features, such as OCC_MANAGER (individuals employed as a manager), OCC_SKILL5 (individuals employed in jobs commensurate with secondary or low level vocational training), and OCC_SKILL4 or UMEMPLOYED (individuals in the labor force and unemployed). This is unsurprising as employment is generally related to educational attainment and its associated lifestyle patterns.^{21,22}

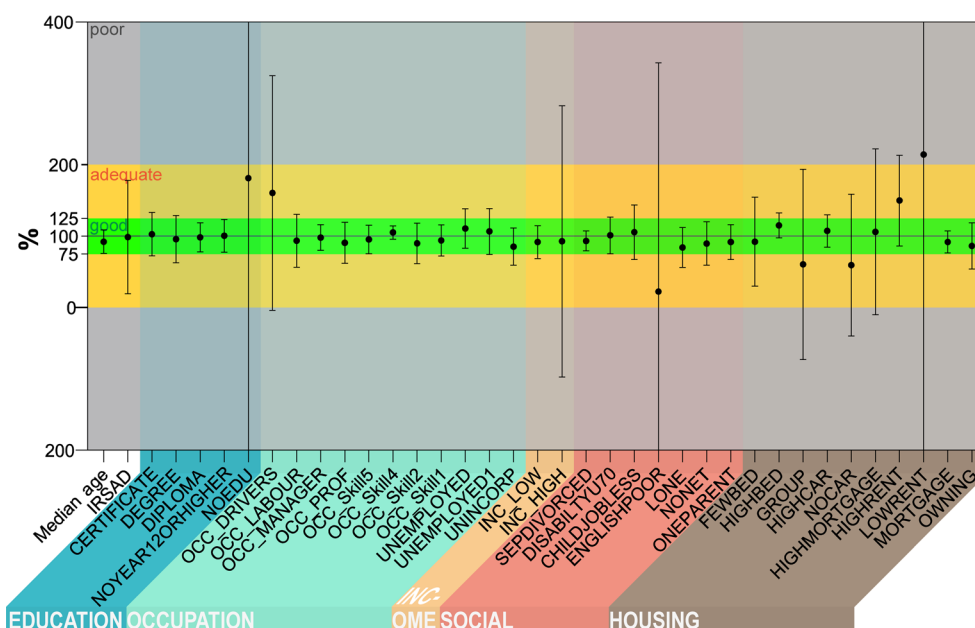


Figure 3. Model precision (mean) and accuracy (\pm SD) for the prediction of socioeconomic features in the test data set. Values were determined by dividing the error of model predictions by census results. Socioeconomic variables other than median age and IRSAD have been grouped into categories to aid interpretation. Refer to Table S1 for feature definitions.

Features relating to social structure were adequately predicted, apart from good prediction for SEPDIORCED (separated or divorced individuals) and poor prediction for ENGLISHPOOR (poor or no English speaking ability). This is supported by studies which show links between social support and diet or health outcomes in many demographics.^{23,24}

Housing features were the most poorly predicted category in this study. This trend is corroborated by environmental studies, which suggest that housing features are a relatively minor determinant of health,²⁵ especially when compared to other factors, such as level of social support, which are better predictors of well being than housing features.²⁶

Among the poorly predicted variables were NOEDU, ENGLISHPOOR, GROUP (homes with multiple households), HIGHRENT (homes paying > \$470 AUD/week in rent), and LOWRENT (homes paying < \$215 AUD/week in rent). This was likely due to the presence of zero and near-zero entries for these variables in the training data set (Table S6). Prediction for features such as LOWRENT and HIGHRENT are also likely to be further confounded by the effect of differences in purchasing power from region to region (e.g., urban vs rural areas). In contrast, NOCAR (homes without a car), HIGHMORTGAGE (homes paying more than \$2800 AUD/month in mortgage), INC_HIGH (individuals with household income in top two deciles), and OCC_DRIVERS (machinery operators and drivers) were predicted poorly despite an absence of zero or near-zero values. These results may therefore indicate that consumption patterns of the measured biomarkers are not as closely associated with these socioeconomic groups. For example, poor prediction performance for INC_HIGH, especially when compared to INC_LOW (individuals with household income in bottom two deciles), may be explained by higher income enabling a wider variety of lifestyle and consumption patterns. Low income, in contrast, is more commonly associated with restrictions in lifestyle and consumption patterns, resulting in relatively more convergent lifestyle patterns. Poor prediction performance for

OCC_DRIVERS may be a result of this category encompassing a wide variety of occupations (e.g., drivers and machinery operators) and hence lifestyles.

Interestingly, with the exception of LOWRENT, all features generally maintained average precision within 0% and 200% (Figure 3), and model predictions were generally randomly distributed around the actual census value of each variable (Figure 1B, Figure 3), with the prediction falling within 75% and 125% for most features. This suggests that while the model may not be precise for individual wastewater samples, predictions of multiple locations (e.g., different areas within a larger city) may be more useful in future applications of this model. Future implementations of similar models would benefit from focusing on population features relevant to more than a small minority of the population and features that can be enumerated accurately. Where possible, multiple samples should be taken across consecutive days or different locations to enhance the overall accuracy of predictions.

The wastewater-based socioeconomic analysis presented here was limited by the quality of calibration (census) data and wastewater data. Future models must be developed using high-quality survey data specific for the WWTP catchments being sampled, which can be achieved using geospatial software.⁷ Furthermore, not all socioeconomic features are expected to be linked to specific chemical consumption patterns (e.g., housing), especially at a population scale. Lastly, the present study employed a specific panel of compounds (Table S2), which were not intended to extensively cover multiple facets of lifestyle and consumption behaviors. Nontargeted mass spectrometry methods²⁷ for biomarker discovery may identify chemical entities that better define and predict population features. Nevertheless, the overall satisfactory performance of the model outside of its training set suggests that the trends observed in the model may be generalizable to larger populations. Models such as the one presented here could be replicated in other nations where some population descriptor(s) can be ascribed to populations served by

WWTPs. This approach could be used to complement population surveys and population censuses by providing an objective estimate of socioeconomic information in areas where such information is infrequently collected, costly to obtain, or rapidly changing.

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.estlett.0c00392>.

Training data set (Table S4) and Test data set (Table S5), (XLSX)

Definitions for socioeconomic index for area (SEIFA) features used in the present study (Table S1), list of biomarkers used in the present study (Table S2), sampling characteristics for WWTPs in the present study (Table S3), summary of socioeconomic variables with entries as zero or under 1% (Table S6), variance in data set explained by model (Figure S1), and WWTP populations featured in the present study (Figure S2) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Saer Samanipour – Queensland Alliance for Environmental Health Sciences, The University of Queensland, Brisbane, QLD 4102, Australia; Norwegian Institute for Water Research, 0349 Oslo, Norway; Van't Hoff Institute for Molecular Sciences, University of Amsterdam, Amsterdam, The Netherlands; orcid.org/0000-0001-8270-6979; Email: s.samanipour@uva.nl

Authors

Phil M. Choi – Queensland Alliance for Environmental Health Sciences, The University of Queensland, Brisbane, QLD 4102, Australia; orcid.org/0000-0002-0535-8197

Jake W. O'Brien – Queensland Alliance for Environmental Health Sciences, The University of Queensland, Brisbane, QLD 4102, Australia; orcid.org/0000-0001-9336-9656

Ben J. Tschärke – Queensland Alliance for Environmental Health Sciences, The University of Queensland, Brisbane, QLD 4102, Australia; orcid.org/0000-0002-3292-3534

Jochen F. Mueller – Queensland Alliance for Environmental Health Sciences, The University of Queensland, Brisbane, QLD 4102, Australia

Kevin V. Thomas – Queensland Alliance for Environmental Health Sciences, The University of Queensland, Brisbane, QLD 4102, Australia; orcid.org/0000-0002-2155-100X

Complete contact information is available at: <https://pubs.acs.org/doi/10.1021/acs.estlett.0c00392>

Funding

The Queensland Alliance for Environmental Health Sciences, The University of Queensland gratefully acknowledges the financial support of the Queensland Department of Health. This project was supported by an Australian Research Council Linkage Project (LP150100364).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors are grateful to the relevant wastewater treatment plant workers who contributed samples towards this study and the anonymous reviewers who reviewed this manuscript.

■ REFERENCES

- (1) Various Authors Valuing the Australian Census; Lateral Economics: 27 August 2019; p 8.
- (2) The Economist Group Limited Costing the Count. <https://www.economist.com/international/2011/06/02/costing-the-count> (accessed 1/24/2020).
- (3) Choi, P. M.; Tschärke, B.; Samanipour, S.; Hall, W. D.; Gartner, C. E.; Mueller, J. F.; Thomas, K. V.; O'Brien, J. W. Social, demographic, and economic correlates of food and chemical consumption measured by wastewater-based epidemiology. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 21864–21873.
- (4) Ort, C.; Lawrence, M. G.; Reungtoat, J.; Mueller, J. F. Sampling for PPCPs in Wastewater Systems: Comparison of Different Sampling Modes and Optimization Strategies. *Environ. Sci. Technol.* **2010**, *44*, 6289–6296.
- (5) Ort, C.; Lawrence, M. G.; Rieckermann, J.; Joss, A. Sampling for Pharmaceuticals and Personal Care Products (PPCPs) and Illicit Drugs in Wastewater Systems: Are Your Conclusions Valid? A Critical Review. *Environ. Sci. Technol.* **2010**, *44*, 6024–6035.
- (6) van Nuijs, A. L. N.; Castiglioni, S.; Tarcomnicu, I.; Postigo, C.; de Alda, M. L.; Neels, H.; Zuccato, E.; Barcelo, D.; Covaci, A. Illicit drug consumption estimations derived from wastewater analysis: A critical review. *Sci. Total Environ.* **2011**, *409*, 3564–3577.
- (7) Tschärke, B. J.; O'Brien, J. W.; Ort, C.; Grant, S.; Gerber, C.; Bade, R.; Thai, P. K.; Thomas, K. V.; Mueller, J. F. Harnessing the Power of the Census: Characterizing Wastewater Treatment Plant Catchment Populations for Wastewater-Based Epidemiology. *Environ. Sci. Technol.* **2019**, *53*, 10303–10311.
- (8) Been, F.; Bijlsma, L.; Benaglia, L.; Berset, J.-D.; Botero-Coy, A. M.; Castiglioni, S.; Kraus, L.; Zobel, F.; Schaub, M. P.; Bücheli, A.; Hernández, F.; Delémont, O.; Esseiva, P.; Ort, C. Assessing geographical differences in illicit drug consumption—A comparison of results from epidemiological and wastewater data in Germany and Switzerland. *Drug Alcohol Depend.* **2016**, *161*, 189–199.
- (9) O'Brien, J. W.; Thai, P. K.; Eaglesham, G.; Ort, C.; Scheidegger, A.; Carter, S.; Lai, F. Y.; Mueller, J. F. A model to estimate the population contributing to the wastewater using samples collected on census day. *Environ. Sci. Technol.* **2014**, *48*, 517–25.
- (10) Zhang, Y.; Duan, L.; Wang, B.; Du, Y.; Cagnetta, G.; Huang, J.; Blaney, L.; Yu, G. Wastewater-based epidemiology in Beijing, China: Prevalence of antibiotic use in flu season and association of pharmaceuticals and personal care products with socioeconomic characteristics. *Environ. Int.* **2019**, *125*, 152–160.
- (11) Thomaidis, N. S.; Gago-Ferrero, P.; Ort, C.; Maragou, N. C.; Alygizakis, N. A.; Borova, V. L.; Dasenaki, M. E. Reflection of Socioeconomic Changes in Wastewater: Licit and Illicit Drug Use Patterns. *Environ. Sci. Technol.* **2016**, *50*, 10065–10072.
- (12) O'Brien, J. W.; Grant, S.; Banks, A. P. W.; Bruno, R.; Carter, S.; Choi, P. M.; Covaci, A.; Crosbie, N. D.; Gartner, C.; Hall, W.; Jiang, G.; Kaserzon, S.; Kirkbride, K. P.; Lai, F. Y.; Mackie, R.; Marshall, J.; Ort, C.; Paxman, C.; Prichard, J.; Thai, P.; Thomas, K. V.; Tschärke, B.; Mueller, J. F. A National Wastewater Monitoring Program for a better understanding of public health: A case study using the Australian Census. *Environ. Int.* **2019**, *122*, 400–411.
- (13) O'Brien, J. W.; Grant, S.; Mueller, J. F.; Tschärke, B. J.; Gerber, C.; White, J. National Wastewater Drug Monitoring Program – Report 1; The University of Queensland and University of South Australia: Australian Criminal Intelligence Commission (ACIC), March 2017; p 64.
- (14) Australian Bureau of Statistics (ABS). *Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA)*; 2033.0.55.001; Commonwealth of Australia, 2016.

- (15) de Jong, S. SIMPLS: An alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 251–263.
- (16) van den Berg, R. A.; Hoefsloot, H. C. J.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **2006**, *7*, 142–142.
- (17) Brereton, R. G. *Applied Chemometrics for Scientists*; John Wiley & Sons: West Sussex, England, 2007.
- (18) *Introduced Random Error*. Australian Bureau of Statistics (ABS). <https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/2901.0Chapter38202016> (accessed 28/7/2019).
- (19) Florence, M. D.; Asbridge, M.; Veugelers, P. J. Diet Quality and Academic Performance. *J. School Health* **2008**, *78*, 209–215.
- (20) Backholer, K.; Spencer, E.; Gearon, E.; Magliano, D. J.; McNaughton, S. A.; Shaw, J. E.; Peeters, A. The association between socio-economic position and diet quality in Australian adults. *Public Health Nutrition* **2016**, *19*, 477–485.
- (21) Nocon, M.; Keil, T.; Willich, S. N. Education, income, occupational status and health risk behaviour. *Journal of Public Health* **2007**, *15*, 401–405.
- (22) Andersen, R.; Van De Werfhorst, H. G. Education and occupational status in 14 countries: the role of educational institutions and labour market coordination. *British Journal of Sociology* **2010**, *61*, 336–355.
- (23) Umberson, D.; Karas Montez, J. Social Relationships and Health: A Flashpoint for Health Policy. *Journal of Health and Social Behavior* **2010**, *51*, S54–S66.
- (24) Pieroth, R.; Rigassio Radler, D.; Guenther, P. M.; Brewster, P. J.; Marcus, A. The Relationship between Social Support and Diet Quality in Middle-Aged and Older Adults in the United States. *J. Acad. Nutr. Diet.* **2017**, *117*, 1272–1278.
- (25) Shaw, M. *Annu. Rev. Public Health* **2004**, *25*, 397–418.
- (26) Wright, P. A.; Kloos, B. Housing environment and mental health outcomes: A levels of analysis perspective. *Journal of Environmental Psychology* **2007**, *27*, 79–89.
- (27) Samanipour, S.; Martin, J. W.; Lamoree, M. H.; Reid, M. J.; Thomas, K. V. Letter to the Editor: Optimism for Nontarget Analysis in Environmental Chemistry. *Environ. Sci. Technol.* **2019**, *53* (10), 5529–5530.