nature sustainability

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41893-018-0153-6

In the format provided by the authors and unedited.

Social-media data for urban sustainability

Rositsa T. Ilieva[1,2]* and Timon McPhearson[1,3,4]

[1]Urban Systems Lab, The New School, New York, NY, USA. [2]CUNY Urban Food Policy Institute, City University of New York, Graduate School of Public Health and Health Policy, New York, NY, USA. [3]Cary Institute of Ecosystem Studies, Millbrook, NY, USA. [4]Stockholm Resilience Centre, Stockholm University, Stockholm, Sweden. *e-mail: ilievar@newschool.edu

Appendix A – Research Methods

We conducted a literature review of scientific publications addressing the emerging opportunities of using "Big Data" from social media for sustainability research as well as the crosscutting challenges, and potential solutions, to social media data (SMD) research that sustainability scholars and decision-makers are likely to face. We focused our discussion on the implications that such challenges have for urban sustainability research and cities more specifically. The review uses replicable methods to identify, analyze, and critically appraise all relevant research on a given topic (Tricco et al., 2009). The effective application of SMD to the study of urban environmental problems, human-nature relationships, and sustainable urban development is a novel area of research where still a very limited number of scholarly contributions exists. To advance knowledge in this direction, it is therefore necessary to learn from the broader discussion about the promises and perils of SMD as a means for augmenting sustainability research in the social, natural, and design fields. The publications reviewed were thus selected from scientific literature specialized in the fields of urban planning and design, applied geography, geographic information systems (GIS), transportation planning, and urban ecology as well as in the fields of information technologies, artificial intelligence, "Big Data," marketing, digital social sciences, and telecommunications.

1. Search strategy

The search strategy for this paper was designed with the aim to collect recent evidence of the application of SMD to the investigation of urban sustainability questions, as well as evidence of key challenges to the integration of SMD in scientific research and potential strategies to overcome them (**Figure 1**). The two thematic streams were kept distinct to afford the identification of the most relevant contributions pertaining to each subtopic, which would otherwise remain undetected in a combined search (e.g., imposing sustainability focus together within a focus on challenges).

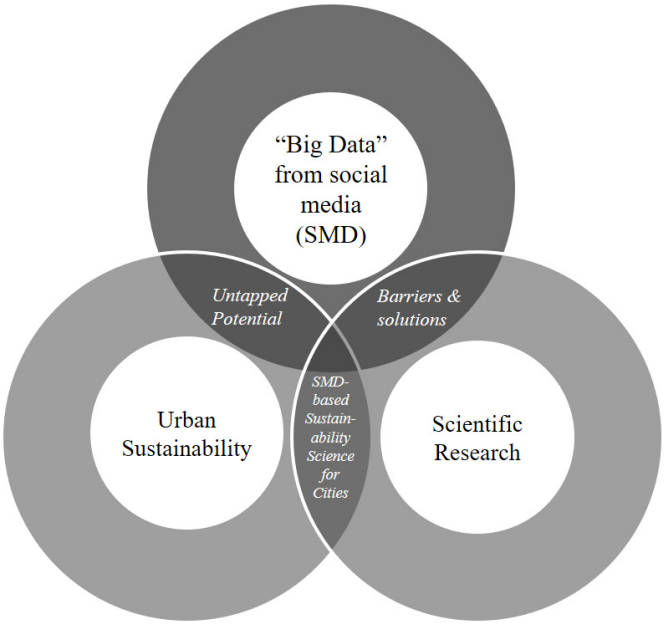

Figure 1 Overview of the search strategy and the intersection of SMD with opportunities for urban sustainability research as well as challenges to integrate SMD in scientific research more broadly.

Articles were retrieved through a database search, using a predefined set of keywords combined through Boolean operators, in three major online databases: Web of Science, ProQuest, and EBSCO Host. The initial set of publications was later supplemented with articles identified through a snowball search technique which involved screening the lists of references of relevant papers to identify useful records not detected through the database search. To select the final sample of papers for the review, papers were analyzed thematically through a hierarchical, cascade-like procedure, starting with screening the papers' titles, abstracts, and eventually full texts (only for the papers most pertinent for this review). To further consolidate the final sample of publications, after screening all records, based on the inclusion and exclusion criteria described below, a final selection step consisted of identifying groups of papers which presented the same conclusions or greatly overlapped; in these instances, the most recent publication was kept. This narrowed the final sample down to 105 records which were included in the final review (**Figure 2**).

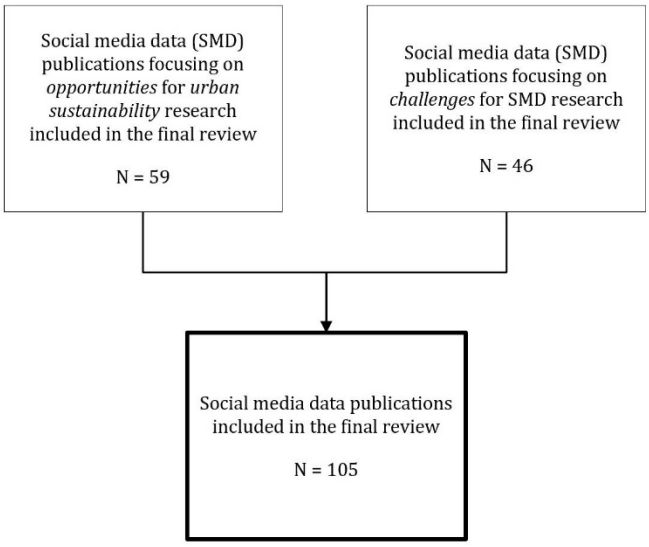

Figure 2 Final sample of articles included in the review.

1.1. Evidence on emerging opportunities of using SMD for urban sustainability research

Keyword searches included terms related to SMD (social media data, geolocated social media, geolocated social networks, location based social networks, location based, Big Data and social media, neogeography, Flickr, Twitter, Instagram, Foursquare) and at least one other term pertaining to urban spaces or activities (city, cities, urban, park, green space, recreation, tourism, transport*, mobility, walk*, visit*, retail, housing) and sustainability (sustainab*, unsustainab*, resilien*, urban ecology, environmental, ecolog*, global warming, climate change, carbon dioxide, heat island, ecosystem services, extreme weather, flooding, vulnerab*, SDGs, New Urban Agenda, 2030 Agenda).

The initial search yielded 575 articles of which 82 were excluded because of duplication, and 395 because of one or more of the exclusion criteria established or lack of access to the full-text records (**Figure 3**). In total, 493 titles and abstracts were reviewed, of which 98 met the inclusion criteria for this review. In addition, 2 records cited in the reference lists of relevant articles were retrieved and added to the data set at the stage of full-texts review. The subsequently included articles were subjected to the same inclusion and exclusion criteria used to screen the set of records at the beginning of the search process. The so-identified subsample was further narrowed by leaving only unique contributions that were not repeating findings presented in other publications in the sample. The final subsample for the emerging opportunities for SMD-based urban sustainability research was of 59 publications and served as a basis for the findings reported in Section 1 of the Review article.

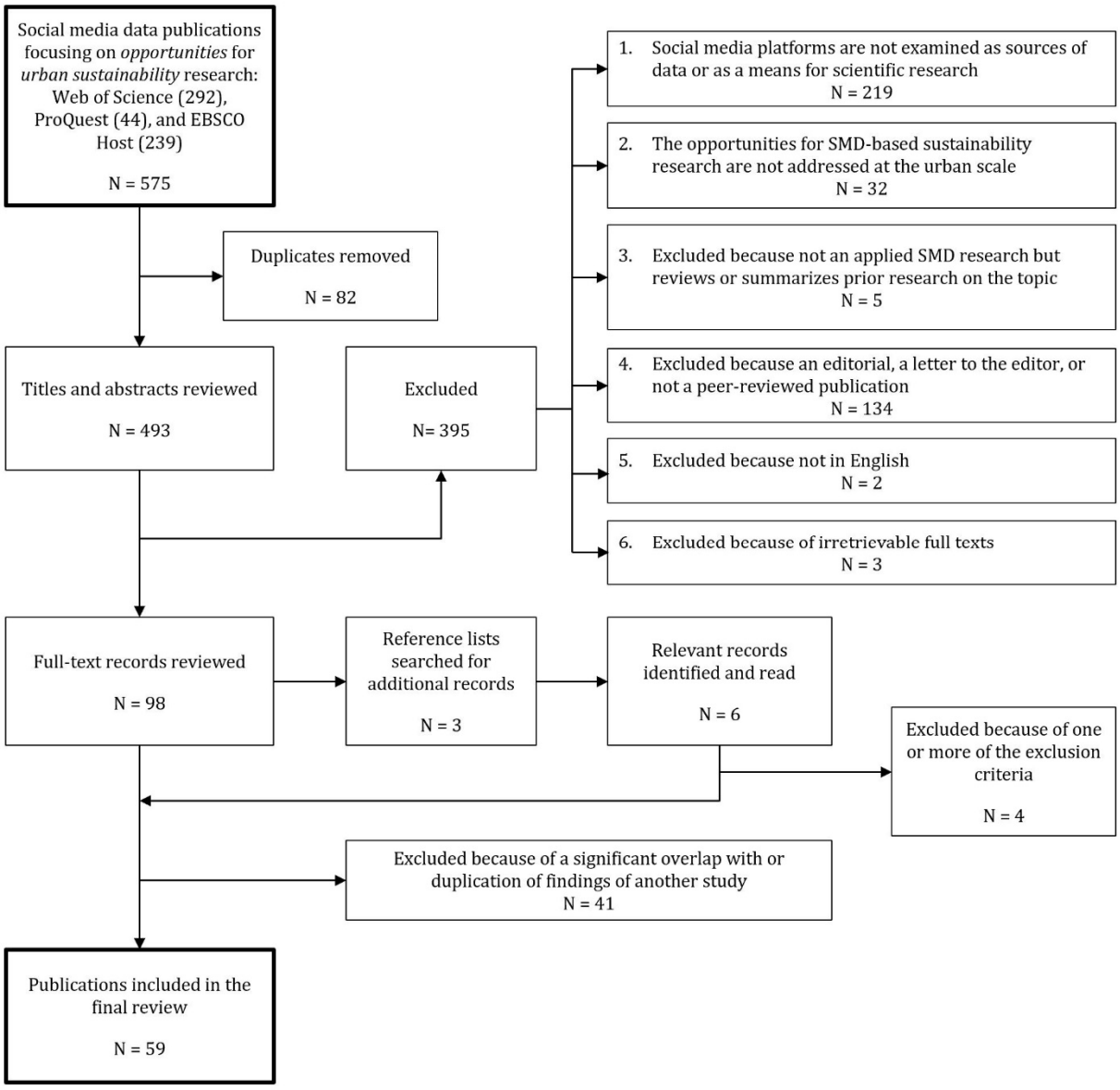

Figure 3 Bibliographic search and inclusion/exclusion process for publications focusing on the promise of SMD for sustainability research.

1.2. Evidence on crosscutting challenges to SMD-based research and potential solutions

Keyword searches focusing on SMD and its application to urban sustainability research included terms related to SMD (social media data, geolocated social media, geolocated social networks, location based social networks, location based, Big Data and social media, neogeography, Flickr, Twitter, Instagram, Foursquare) and at least one other term pertaining to challenges (bias*, challenge*, limit*, weakness*, shortcoming*, pitfall*, flaw*, constraint*, drawback*).

The initial search yielded 1016 articles of which 408 were excluded because of duplication and 523 because of one or more of the exclusion criteria established or irretrievable full-text records (**Figure 4**). In total, 608 titles and abstracts were reviewed, of which 85 met the inclusion criteria for this review. In addition, 3

records cited in the reference lists of relevant articles were retrieved and added to the data set at the stage of full-texts review. The subsequently included articles were subjected to the same inclusion and exclusion criteria used to screen the set of records at the beginning of the search process. The so-identified subsample was further narrowed by leaving only unique contributions that were not repeating findings presented in other publications in the sample. The final subsample for the SMD challenges and solutions topic was of 46 publications and served as a basis for the findings reported in Section 2 of the Review article.

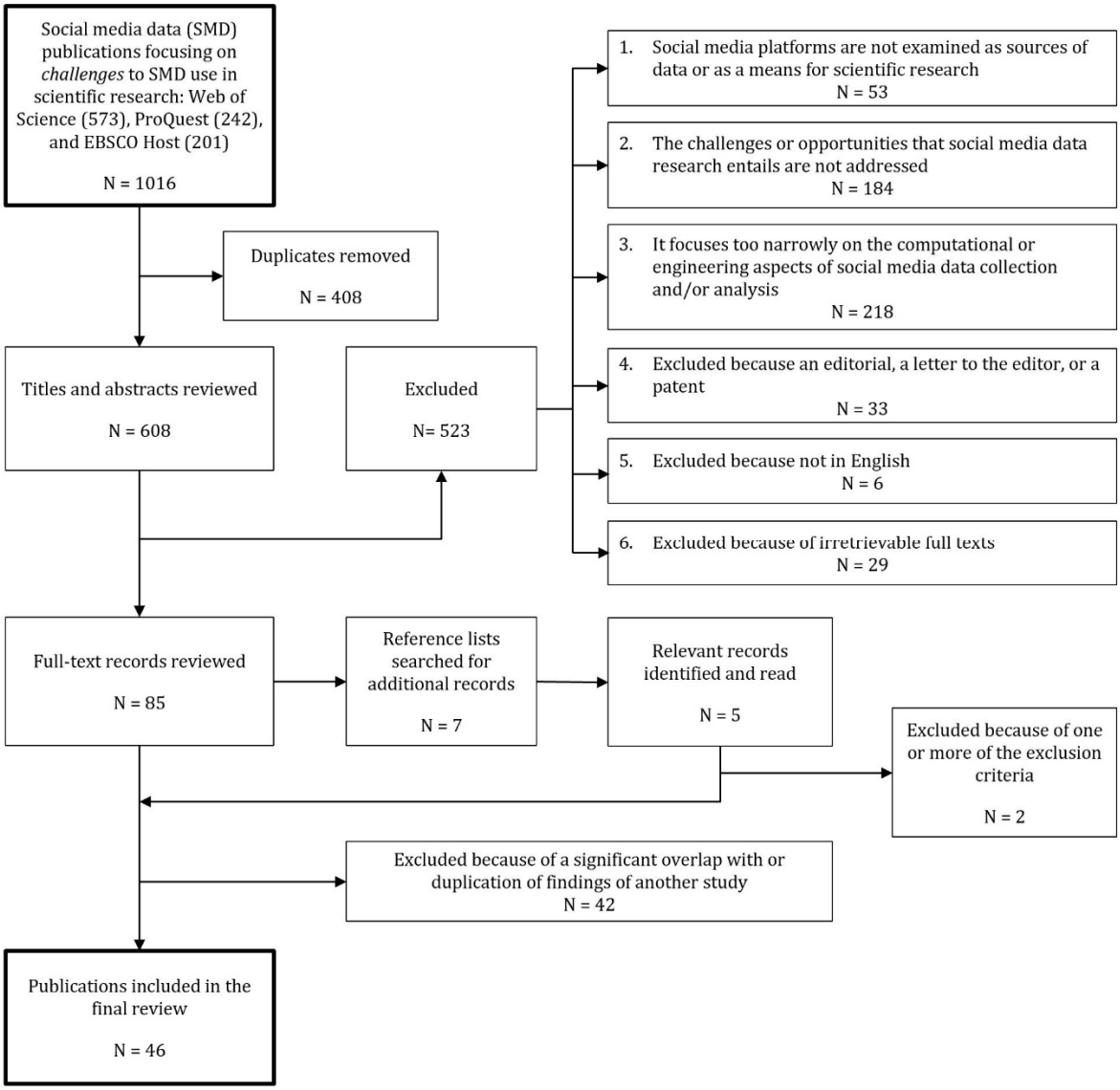

Figure 4 Bibliographic search and inclusion/exclusion process for publications focusing on the challenges of SMD for research and potential strategies to overcome them.

2. Inclusion and exclusion criteria

The selection of relevant publications for the Review was guided by a set of inclusion and exclusion criteria outlined at the beginning of the bibliographic search process (**Table 1** and **Table 2**). The types of scientific publications included were peer-reviewed academic journals articles, book chapters, and, in some cases if

particularly pertinent to this review and only if no comparable journal publication existed, scientific papers published in peer-reviewed conference proceedings. The search included records published through December 2017.

2.1. Evidence on emerging opportunities of using SMD for urban sustainability research

To be included in Section 1 of the review, focusing on opportunities of using SMD in urban sustainability research, a publication had to focus on SMD as a means for scientific research, discuss the opportunities for using SMD to address urban questions, relate to key sustainability challenges, and report on an applied research study. Records were excluded from the review if they did not focus on SMD as source of "Big Data" but discussed uses of social media platforms per se (e.g., the use of Facebook in education), were literature reviews, were not peer-reviewed publications, or were not in English.

To be considered a paper using SMD to research human-environment interactions, an article had to deploy "Big Data" from social media to advance knowledge in one or more domains of human-environment interactions science (Stern, 1993) such as: human causes of environmental change (including also indirect causes such as changes of human values and institutions), the effects of environmental change on what people value (open spaces, red-list plant and animal species, natural resources, agricultural productivity), or actions people take to anticipate environmental damage and preserve environmental values (projects, programs, policies, or management plans). The subdomain of human-nature relationships (Flint et al., 2013) with specific focus on people's direct and personal experiences with nature and urban green spaces (Kabisch et al., 2015) and parks was considered as part of this inclusion criterion.

To be selected as a paper using SMD to research social behavior in cities, a paper had to employ a SMD-based analysis of human activity patterns, such as working, studying, shopping, eating, travelling, recreation, and socializing, in cities (Chapin, 1974; Bechtel, 1997). The emphasis was thus placed on social spatial behavior, which can be broadly understood as comprised of "consciously or subconsciously directed life process that result change location through time" (Golledge and Stimson, 1997, p. 155).

Table 1 Inclusion and exclusion criteria for publications focusing on the promise of SMD for sustainability research.

Inclusion	Exclusion
Discusses the new opportunities offered by SMD for one or more spheres of urban sustainability research	Social media platforms are not examined as sources of "Big Data" for scientific research
Uses SMD to research social behavior in cities	The opportunities of using SMD to address sustainable development questions are not discussed
Uses SMD to research human-environment interactions	Does not focus on cities or urban areas
Peer-reviewed	Not peer-reviewed
Published through December 2017	Not in English

2.2. Evidence on crosscutting challenges to SMD-based research and potential solutions

To be included in Section 2 of the review, discussing crosscutting challenges to the use of SMD in scientific research, a publication had to explicitly discuss current limitations of SMD and/or strategies to address current limitations of different aspects of SMD-based research. Records were excluded from the review if they did not focus on SMD as source of "Big Data" but discussed uses of social media platforms per se (e.g., the use of Facebook in education), focused too narrowly on the computational or engineering aspects of SMD research, were not peer-reviewed, were editorials or letters to the editor, or were not in English. Markedly technical or engineering studies were included only in the cases in which they addressed a specific challenge to the use of SMD in research and/or its solution.

Table 2 Inclusion and exclusion criteria for publications focusing on the challenges of SMD for research and potential strategies to overcome them.

Inclusion	Exclusion
Discusses challenges or barriers to the use of SMD for scientific research	Social media platforms are not examined as sources of "Big Data" for scientific research
Discusses potential solutions to key challenge/s to the use of SMD in scientific research	The challenges of SMD for scientific research are not discussed
Peer-reviewed	Focuses narrowly on the computational or engineering aspects of SMD collection and/or analysis
Published through December 2017	Not peer-reviewed
	Not in English

3. Data coding

This review is the outcome of a structured analysis and synthesis of emergent opportunities to use "Big Data" from social media in different fields of sustainability research, and related professional domains, as well as crosscutting challenges. To this end, two distinct, yet interrelated, templates of key themes that could guide data extraction were designed based on influential studies in the field and inductive analysis of the papers in the final sample.

Evidence on emerging opportunities of using SMD for urban sustainability research

A conceptual framework discerning five different domains of sustainable urban development – environmental sustainability, public health, social equity, mobility, and economic vitality – was adopted to thematically organize the lessons drawn from each SMD publication. This allowed for a more integrated perspective on the promises of SMD for different spheres of sustainability research and helped overcome some of the limitations of in-depth accounts on separate applications, which render only a piecemeal answer to the central research question raised in the paper. Papers selected for the final subsample focusing on the use of SMD for the investigation of urban sustainability questions were, therefore, examined with respect to these five sustainability dimensions. Data extraction was guided by thematic analysis and classification of each paper's focus and use of SMD, which were illustrated in the paper's methods and results sections.

The choice of the five normative dimensions in the conceptual framework for urban sustainability were based on the three core tenets of the Brundtland Report (1987), which advanced the three-fold objective for environmental, social, and economic sustainability, and the model more recent extensions through scholarship in the public health literature and the 2030 Agenda for Sustainable Development. We are also sympathetic with Raworth's (2012) framework of "safe and just (operational) space" suggesting that society move to a space whereby deprivations (e.g., hunger, poverty) are overcome and, at the same time, a safe distance is kept from resource limits and planetary boundaries. By including a separate dimension on transportation, we also build on the perspective suggested by O'Neill and colleagues (2018) who posit that in order to operationalize "strong sustainability," or the preservation of *natural* and *social capital* in tandem, we need to include *systems of provision*, which mediate and shape the communication between people and nature.

Evidence on crosscutting challenges to SMD-based research and potential solutions

To be considered relevant for the development of the data extraction template on challenges to the deployment of SMD in academic research, a study had to explicitly focus on this topic. Studies where a

limitation or an advantage of SMD were mentioned only in passing or addressed only in a subsection of a paper were thus not considered. The final data extraction template featured 6 broad categories of potential sources of bias or challenges noted by authors in SMD research – representativeness of sample population, location of human behavior, analysis of sentiments and opinions, temporality of content, ethical dilemmas, and reliability of data analysis. During the in-depth review of the final set of publications, relevant subcategories were also identified, thereby refining the initial taxonomy of challenges to SMD-based sustainability research and their potential solutions (see Appendix C, Table 1). Constant comparisons between the publications examined afforded the iterative assessment of the differences and similarities in the definition of the opportunities, biases, and possible solutions across different studies. The critical appraisal of each study was guided by the questions on the provenience, purpose, and reliability of each source, outlined by Stewart and Kamins (1993) for secondary data research.

4. Limitations

While effort was made to design a comprehensive and thoroughly documented search and analytical process, the present review has limitations. The search included only peer-reviewed literature written in English, thus potentially omitting recent academic literature on urban SMD published in non-English scientific journals or in non-peer-reviewed academic periodicals. Additionally, to maintain a reasonable balance between the breadth and depth of the review, the selection strategy used for the review purposefully excluded articles which focused exclusively on the computational aspects of SMD. This may have led to the omission of ICT articles with implications for urban sustainability research or the solutions to identified SMD challenges. Finally, the appraisal and classification of the literature into emerging opportunity areas for sustainability research was shaped by a specific view of sustainability, grounded in the Brundtland Report and the 2030 Agenda for Sustainable Development. Other scholars may have discerned different crosscutting themes or emphasized different associations between SMD applications and emergent opportunities for sustainability research. In fact, budding applications of SMD to city planning often have bearing on more than one sustainability domain.

References cited

Bechtel, R. B. (1997). *Environment and behavior: An introduction*. New York, NY: Sage.

Chapin, F. S. (1974). *Human activity patterns in the city: Things people do in time and in space*. New York, NY: Wiley-Interscience.

Flint, C. G., Kunze, I., Muhar, A., Yoshida, Y., & Penker, M. (2013). Exploring empirical typologies of human-nature relationships and linkages to the ecosystem services concept. *Landscape and Urban Planning*, *120*, 208–217.

Golledge, R.G. and Stimson, R.J. (1997): Spatial Behavior: A Geographic Perspective. New York, NY: Guilford Press,

Kabisch, N., Qureshi, S., & Haase, D. (2015). Human–environment interactions in urban green spaces - A systematic review of contemporary issues and prospects for future research. *Environmental Impact Assessment Review*, *50*, 25–34.

O'Neill, D. W., Fanning, A. L., Lamb, W. F., & Steinberger, J. K. (2018). A good life for all within planetary boundaries. *Nature Sustainability*, *1*(2), 88–95.

Raworth, K. (2012). A safe and just space for humanity: can we live within the doughnut. *Oxfam Policy and Practice: Climate Change and Resilience*, *8*(1), 1–26.

Stern, P. C. (1993). A second environmental science: Human-environmental interactions. *Science*, *260*(5116), 1897–1899.

Stewart, D. W., & Kamins, M. A. (1993). *Secondary research: Information sources and methods* (Vol. 4.). Newbury Park, CA: SAGE Publications.

Tricco, A. C., Tetzlaff, J., & Moher, D. (2011). The art and science of knowledge synthesis. *Journal of Clinical Epidemiology*, *64*(1), 11–20.

Appendix B – Results: Supplementary Information

In total, 105 articles met our inclusion criteria for this research. Figure 1 summarizes the distribution of the papers examined per typology of social media platform, while **Figure 2** shows the share of papers retrieved per urban sustainability topic. An overview of the findings on the chief challenges and sources of bias in SMD research, as well as the possible strategies to overcome such challenges, are presented in Appendix C, Table 1.

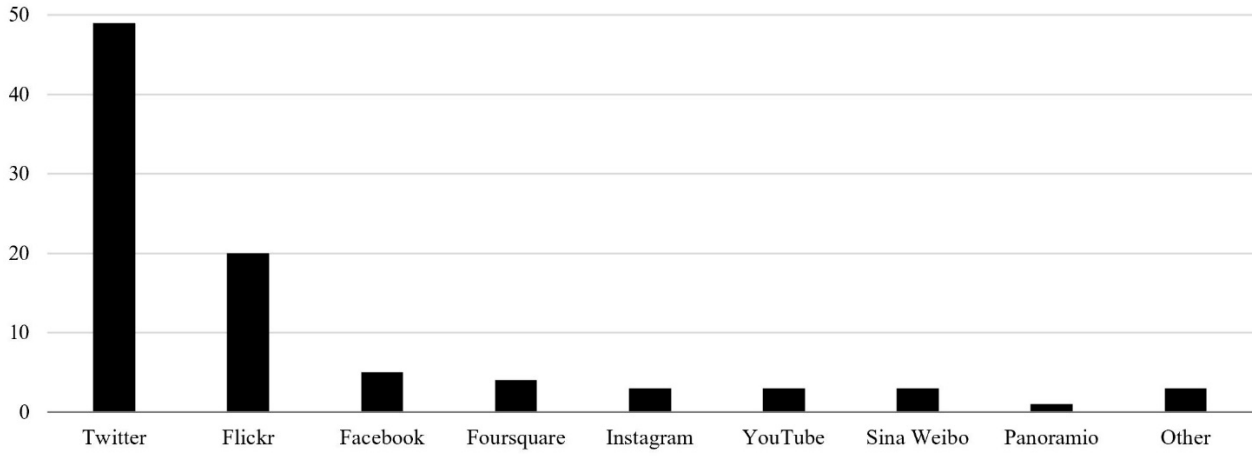

Figure 1 Papers reviewed per social media platform. *Note*: Conceptual and literature review papers addressing SMD in general were not considered for this chart.

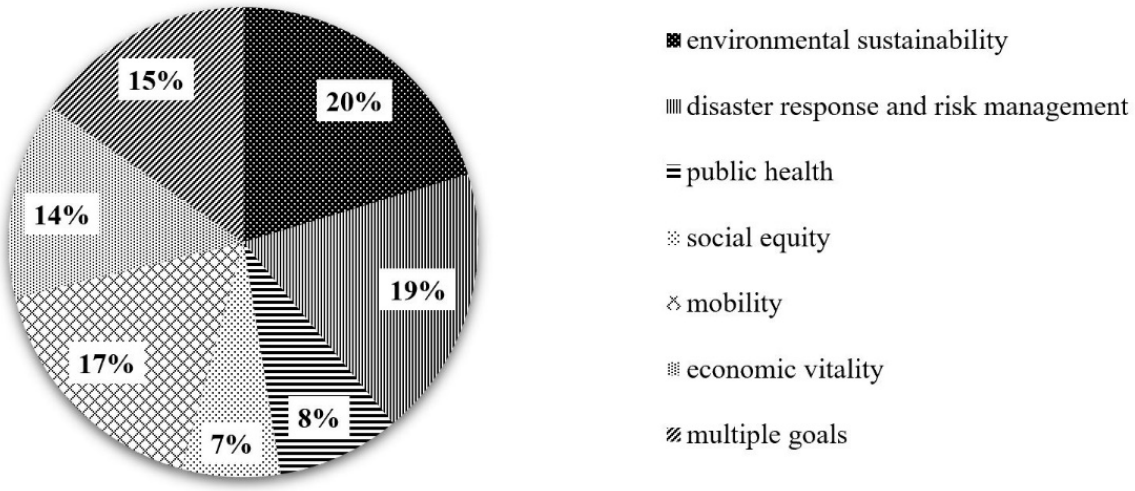

Figure 2 SMD papers per type of urban sustainability goals addressed.

Appendix C – Results: Crosscutting Challenges & Potential Solutions

Table 1. Detailed Overview of Crosscutting Challenges to SMD Research and Potential Solutions

Overarching challenge	Specific challenge	Studies noting challenge	Possible solution	Studies noting solution
Representation of population	**Accounts vs. People**	boyd and Crawford, 2012 Ruths and Pfeffer, 2014	Use multiple sources of data	Kosinski et al., 2015
			Use qualitative research methods	Baym, 2013
	Listeners vs. Active users vs. Very active users	boyd and Crawford, 2012; Giglietto et al., 2012	Use models able to isolate the top users contributing to a specific activity pattern	Hasan and Ukkusuri, 2014
	Photographers vs. Visitors	Levin et al., 2015; Wood et al., 2013	Use local authority censuses to confirm visitor numbers	Levin et al., 2015
			Use total empirical annual visitor user-days	Wood et al., 2013
	Nonhumans in large-scale studies	Ruths and Pfeffer, 2014 Baym, 2013 Crampton et al., 2013	Filters or algorithms for detecting nonhuman accounts	Ruths and Pfeffer, 2014 Crampton et al., 2013
	Proxy population	Kosinski et al., 2015; Ruths and Pfeffer, 2014	Correct platform-specific and proxy population biases	Ruths and Pfeffer, 2014
			Sampling methods decoupled from substantive research questions	Hargittai, 2015
	Missing or skewed demographics	Allan et al., 2015; Baym, 2013; Guerrero et al., 2016; Huang and Park, 2013; Wood et al., 2013	Derive metadata from profile descriptions	Dunkel, 2015 Sloan et al., 2015
			Demographic collators	boyd and Crawford, 2012; Kosinski et al., 2015; Pettit, 2011

Overarching challenge	Specific challenge	Studies noting challenge	Possible solution	Studies noting solution
			Use face recognition algorithms	Jiang et al., 2015
	Voluntary participation	Goodspeed, 2013; Hargittai, 2015	New statistical methods for generalization from unrepresentative samples	Goodspeed, 2013
Representation of activities	**Activities are differently suited to photos**	Allan et al., 2015; Wood et al., 2013	Correspondence between photo-user-days and surveyed visits	Keeler et al., 2015; Wood et al., 2013
	Events are differently suited to comments	Diaz et al., 2016	n/d	n/d
	Missing activities	Hasan and Ukkusuri, 2014	Activity pattern recognition model to predict missing activities	Hasan and Ukkusuri, 2014
	Sayings vs. doings	Baym, 2013; Goodspeed, 2013; Hargittai, 2015	Rich record on both sayings and doings	Goodspeed, 2013
			Datasets from multiple social network sites	Hargittai, 2015
	Platform-driven *vs.* psychosocial behaviors	Baym, 2013; Goodspeed, 2013; Kosinski et al., 2015; Ruths and Pfeffer, 2014	Untangle psychosocial from platform-driven behavior	Ruths and Pfeffer, 2014 Goodspeed, 2013
	Users' skills	Hargittai, 2015	Make explicit the limitation of each sample used	Hargittai, 2015
	Distortion of human behavior	Ruths and Pfeffer, 2014 Baym, 2013	n/d	n/d
Representation of space	**Ordinary places**	Dunkel, 2015; Wood et al., 2013	Use alternative SMD as proxy for visitation	Wood et al., 2013
	Sharable places	Cranshaw et al., 2012	n/d	n/d
	Images without GPS-based coordinates	Hauff and Houben, 2012	Use image's textual metadata	Serdyukov et al., 2009
			Use the image user's traces on other social	Hauff and Houben, 2012

Overarching challenge	Specific challenge	Studies noting challenge	Possible solution	Studies noting solution
			media platforms	
	Geotags vs. Exact location	Crampton et al., 2013; Han et al., 2015; Leetaru et al., 2013; Schwartz and Halegoua, 2014	Text-based metadata and traditional and fulltext geocoders	Leetaru et al., 2013
			Social network analysis considering the density of retweets	Crampton et al., 2013
			Interpret geotags/user-provided as forms of self-representation	Schwartz and Halegoua, 2014
			Distinguish different cities with the same name and dissimilar semantics	Han et al., 2015; Leetaru et al., 2013
			Borrow search techniques from AI, information retrieval and natural language processing	Naaman, 2011
	Proximity of buildings in urban settings	Crooks et al., 2015	Combine crowdsourced data services and use Latent Dirichlet Allocation LDA to discover topics within geo-located data	Crooks et al., 2015
	Upload location vs. Exact location	Guerrero et al., 2016	Individually analyze images with upload location errors	Guerrero et al., 2016
	Uneven geographies of platform usage	Cranshaw et al., 2012; Quercia and Saez, 2014	Produce results for the whole city and for the top 50 neighborhoods	Quercia and Saez, 2014
	Number of photos vs. Place popularity	Dunkel, 2015	Collapse all photos taken by a single user within a given radius to a single arithmetically centered point	Dunkel, 2015
			Consider new units of analysis such as “regions of attraction” estimating the visiting probability of a place	Kou et al., 2015
	Scarcity of GPS-coded geotags	Crampton et al., 2013; Edwards et al., 2013; Leetaru et al., 2013	n/d	n/d

Overarching challenge	Specific challenge	Studies noting challenge	Possible solution	Studies noting solution
Representation of sentiments	**Platform-side filtration**	boyd and Crawford, 2012	n/d	n/d
	Subjectivity	Chan et al., 2016	Use statistical cluster analysis	Chan et al., 2016
			Use an opinion bias detection model	Kwon and Lee, 2013
			Analyze words out of context	Gore et al., 2015
	Content poverty	Crampton et al., 2013; Edwards et al., 2013; Goodspeed, 2013	Collect contextual information about users' choices, goals, and activities	Goodspeed, 2013
	Sayings vs. Thoughts	Bacallao-Pino, 2014; Goodspeed, 2013; Grieve et al., 2014	Participatory research methods, mixed-methods research design, action research	Goodspeed, 2013
	Purposeful Misinformation	Schoen et al., 2013	n/d	n/d
	Language and culture	Batrinca and Treleaven, 2014; Crampton et al., 2013; Hong et al., 2012; Leetaru et al., 2013	Use social network analysis for level and frequency of connections between individuals	Crampton et al., 2013
	Sarcasm and metaphors	Widener and Li, 2014	n/d	n/d
	Neutral scores	Pettit, 2011	Independent raters score thousands of records, across various categories and from various data sources	Pettit, 2011
Temporality	**Data volatility**	Crampton et al., 2013; Croitoru et al., 2013; Diaz et al., 2016; Pettit, 2011; Widener and Li, 2014	Develop new tools for measuring and monitoring the resilience of social and geographic ties	Croitoru et al., 2013
			Use relatively large sample sizes e.g., >100,000	Pettit, 2011
			Treat SMD as an imperfect continuous panel survey	Diaz et al., 2016

Overarching challenge	Specific challenge	Studies noting challenge	Possible solution	Studies noting solution
			Ensure consistent availability and update of large datasets	Young, 2014
	Short-lived events	Crampton et al., 2013	n/d	n/d
	No start/end times of activity	Hasan and Ukkusuri, 2014	n/d	n/d
	Behavioral norms have temporal nature	Ruths and Pfeffer, 2014	n/d	n/d
Ethics and privacy	**Participants not asked**	Giglietto et al., 2012; Guerrero et al., 2016; Kennedy, 2012; Pettit, 2011; Schwartz and Halegoua, 2014	Assess how SMD research complies with IRB guidelines	Cote, 2013
			Anonymize data. No interaction or communication with the individuals in the sample.	Kosinski et al., 2015
			Use only geo-tagged and time-stamped information without accessing personal details	Frias-Martinez and Frias-Martinez, 2014
	Amateur researchers	Kosinski et al., 2015; Pettit, 2011	n/d	n/d
	Sharing content with third parties	Kennedy, 2012; Sui and Goodchild, 2011	n/d	n/d
	Safely storing data	Batrinca and Treleaven, 2014	Data needs to be secured and ownership and IP issues resolved.	Batrinca and Treleaven, 2014
			Host data-download applications on a restricted access website	Ben-Harush et al., 2012
	Demographic collators	Acquisti and Gross, 2009; Croitoru et al., 2013;	Randomize the serial numbers assigned to new SSNs	Acquisti and Gross, 2009

Overarching challenge	Specific challenge	Studies noting challenge	Possible solution	Studies noting solution
		Crooks et al., 2015	Re-conceptualize privacy	Croitoru et al., 2013
Data access and data quality	**Different providers grant different access to data**	Batrinca and Treleaven, 2014; boyd and Crawford, 2012; Edwards et al., 2013; Giglietto et al., 2012	n/d	n/d
	Price of data	Batrinca and Treleaven, 2014; Kennedy et al., 2013	Qualitative studies of cultures of large-scale, quantitative data	Kennedy et al., 2013
	Proprietary algorithms for public data	Batrinca and Treleaven, 2014; Ruths and Pfeffer, 2014	Identify specific aspects of the behavior of proprietary systems to begin reporting biases	Ruths and Pfeffer, 2014
	Platform-side filtration	Ruths and Pfeffer, 2014	Quantify biases, show results for more than one platform and for time-separated datasets from the same platform	Ruths and Pfeffer, 2014
	Quality of user-generated data	Crampton et al., 2013	Combine user-generated with other data sets	Crampton et al., 2013
			Use metadata to assess data quality	Immonen et al., 2015
Data analysis	**Researcher bias**	boyd and Crawford, 2012	Make explicit the limits of the possible questions and interpretations	boyd and Crawford, 2012
	Incomparability of methods and data	Ruths and Pfeffer, 2014	Transparent research methods	Ruths and Pfeffer, 2014
	Unstructured data	Batrinca and Treleaven, 2014; Hasan and Ukkusuri, 2014; Immonen et al., 2015	Machine learning techniques to find spatio-temporal patterns like topic modeling	Hasan and Ukkusuri, 2014
			Use metadata management to evaluate and manage the quality and trustworthiness of SMD	Immonen et al., 2015

Overarching challenge	Specific challenge	Studies noting challenge	Possible solution	Studies noting solution
	DIY analytic systems	Baym, 2013; Pettit, 2011	n/d	n/d
	Majority bias	Cranshaw et al., 2012	n/d	n/d
	False patterns	Wood et al., 2013	n/d	n/d
	Interoperability	Batrinca and Treleaven, 2014; Croitoru et al., 2013; Crooks et al., 2015; Crosas et al., 2014; Housley et al., 2014	Set uniform units of analysis and data formats across institutions and software	Housley et al., 2014
			Use of more integrated analytical tools with privacy protection	Crosas et al., 2014
			Use a new system integrating heterogeneous social media feeds	Croitoru et al., 2013
	Interdisciplinary work	Crosas et al., 2014; Giglietto et al., 2012; Young, 2014	Research teams combining expertise in geography, computational social sciences, linguistics, and computer science	Young, 2014
	Volume	Croitoru et al., 2013; Crooks et al., 2015; Crosas et al., 2014; Giglietto et al., 2012; Cao et al., 2014; Ediger et al., 2010; Yang et al., 2017	Automatic semantic analysis to support manual coding	Giglietto et al., 2012
			Distributed file storage systems (e.g., Hadoop, Google Drive, Dropbox)	Yang et al., 2017
			Parallel computing (e.g., MapReduce),	Yang et al., 2017
			Data indexing for large-scale spatial queries (e.g., SpatialHadoop, B-tree),	Yang et al., 2017
			Massive social network analysis	Ediger et al., 2010
			Data cube model for location-based social media data analytics	Cao et al., 2014
Validation	**Lack of validation**	Pettit, 2011	Independent raters score thousands of	Pettit, 2011

Overarching challenge	Specific challenge	Studies noting challenge	Possible solution	Studies noting solution
	research		records, across various categories and from various data sources	
			Mixed-methods approaches including ethnographic, statistical, and computational methods	Giglietto et al., 2012
			Validate land-use and social dynamics clustering models with personal interviews and surveys	Cranshaw et al., 2012
			Validate land-use maps derived from geotagged tweets by using real land use data provided by city planning departments and test them in different cities	Frias-Martinez and Frias-Martinez, 2014
			Replicate studies and falsify them	Ruths and Pfeffer, 2014
	Multiple hypotheses testing	Ruths and Pfeffer, 2014	Make visible failed studies. For new methods: compare results to existing methods on the same data. For new social phenomena or methods or classifiers: report performance on two or more distinct data sets one of which was not used during classifier development or design.	Ruths and Pfeffer, 2014
	Reuse of data	Crosas et al., 2014; Gore et al., 2015	Create a framework for sharing, citing, and reusing "Big Data" that supports extensible storage options, allows users to cite subsets of the data with a persistent link and attribution to the data authors, and allow adding metadata.	Crosas et al., 2014