Received: 25 September 2020 | Revised: 20 May 2021 | Accepted: 2 June 2021

DOI: 10.1002/sd.2221

RESEARCH ARTICLE

Sustainable Development | WILEY

Fair weather forecasting? The shortcomings of big data for sustainable development, a case study from Hubballi-Dharwad, India

Andrew Sudmant[1] | Vincent Viguié[2] | Quentin Lepetit[2] | Lucy Oates[1] | Abhijit Datey[3] | Andy Gouldson[1] | David Watling[4]

[1]School of Earth and Environment, University of Leeds, Leeds, UK

[2]Centre International de Recherche sur l'Environnement et le Développement, Paris, France

[3]Department of Energy and Environment, Teri SAS, New Delhi, India

[4]Institute for Transport Studies, University of Leeds, Leeds, UK

Correspondence
Andrew Sudmant, School of Earth and Environment, University of Leeds, Leeds, UK.
Email: a.sudmant@leeds.ac.uk

Funding information
Department for International Development, Grant/Award Number: 113550

Abstract

Sustainable urban mobility is an essential component of sustainable development but requires careful planning in rapidly growing urban areas. This paper investigates the value and limitations of Big Data for evaluating transport policies, plans, and projects in Hubballi-Dharwad, India. Results show how Big Data can enable the outcomes of transport interventions to be evaluated more readily than conventional transport analysis. However, the analysis also found that this data may be less able to detect the impacts of travel behaviours in informal settlements, and the impact of extreme weather events. These potential shortcomings, as well as a lack of transparency around the methodology and data sources used by sources of Big Data, could generate unintended consequences and biases in transport planning. Reflecting on these challenges, and the wider implications for urban governance, we conclude that there is an urgent need for Big Data and other technical advances in urban modelling to be seen as compliments to, rather than substitutes for, wider methods of knowledge generation in urban areas.

KEYWORDS
big data, Google maps, smart cities, sustainable development, sustainable urban mobility, transport

1 | INTRODUCTION

Whether urban growth contributes to solving or aggravating, a wide-range of global challenges will be significantly influenced by urban transport networks. Designed well, transport networks can increase the effective density of urban areas, allowing people to access jobs (and employers to access employees), residents to choose public or non-motorised mobility over their cars, and governments to cost-effectively provide basic services and public amenities (Behbahani, Nazari, Kang, & Litman, 2019; Litman, 2018). Designed poorly, transport networks can choke an urban area's economy and lead to sprawl, congestion, and pollution, making urban dwellers worse off than their rural counterparts.

The cost of congestion in urban areas today, a symptom of transport networks failing to meet demand, is greater than 1% of GDP in European and North American cities, 2–5% of GDP in cities in many Asian and Latin American cities, and as high as 15% of GDP in Beijing, China. While congestion is a global challenge, its costs are most acutely felt in the developing world, making urban transport a critical issue for sustainable development (Cookson & Pishue, 2017; Creutzig & He, 2009; Mao, Yang, Liu, Jianjun, & Jaccard, 2012; Sudmant, Mi, Oates, Tian, & Gouldson, 2020; Sudmant, Verlinghieri, Khreis, & Gouldson, 2020; Too & Earl, 2010; Torres, Ortega, Sudmant, & Gouldson, 2021).

Addressing these challenges will be dependent on the ability of policymakers to rapidly assess the potential for transport

interventions, particularly those that discourage private transport options, and to draw lessons from completed projects. Such assessments, however, face a range of challenges, from overly optimistic modelling processes to problems with data access, to the complexity of modelling urban mobility processes (Gouldson, Sudmant, Khreis, & Papargyropoulou, 2020). Consequently, high-quality ex-post assessments of transport interventions that could yield important insights are rare even in high-income nations (Driscoll, 2014; Flyvbjerg, Skamris Holm, & Buhl, 2003; Nicolaisen & Driscoll, 2014). In urban areas in low-income countries, demand for mobility is rising quickly and public resources face competing demands, addressing this challenge has particular urgency (Cabannes & Lipietz, 2018; Colenbrander et al., 2017).

In this context, tremendous potential is thought to exist from harnessing Big Data: The vast amounts of information coming from mobile phones and other connected devices that are increasingly ubiquitous to our lives. Real-time, geolocated, high frequency, and (in many cases) low-cost applications of Big Data for transport—including Google Maps, Waze, Apple Maps, TomTom, and a host of other services—are already used by billions on a daily basis, ostensibly demonstrating that they are valued by individuals and businesses.

The value of such data from a public policy context, however, does not naturally follow from such services being widely used by individuals and firms (Khan et al., 2020). Big Data sources generally provide a restricted number of variables, requiring assessments to draw inferences with explanatory characteristics (Hu & Jin, 2017). Datasets can be biased, blinding policymakers to the impact of policies on particular populations (Kwan, 2018). The way algorithms capture, sort, clean, and pass on data can alter our understanding of phenomena in ways that policymakers (and sometimes information providers) are not aware of (Zou & Schiebinger, 2018). Modes of governance informed by such data can incentivise city governments to prioritise a narrow set of metrics (Hughes, Giest, & Tozer, 2020) and to discount wider means of urban knowledge generation (Coletta & Kitchin, 2017). And what data is available, for who, and under what circumstances remains a legally and ethically contentious question, with a number of authors reminding us that it would be naive to assume that the interests of private firms automatically align with the interests of the wider public (Albino, Berardi, & Dangelico, 2015; Docherty, Marsden, & Anable, 2018; Wang & Ma, 2021). Questions surrounding the value of Big Data for policymaking thus extend from the specificities of data collection techniques and the ways algorithms are developed to overarching logics and rationalities and their implications for governmentalities (Bissell, 2018; Coletta & Kitchin, 2017; Kitchin, Lauriault, & McArdle, 2015).

Nonetheless, "Smart Cities" relying heavily on Big Data have become a national policy objective in a number of countries worldwide. In India, the Smart Cities Mission was launched in 2015 to support sustainable development through the application of information and communication technologies (Dwevedi, Krishna, & Kumar, 2018). The Smart Cities Mission is focused on cities with a population between 1 and 4 million ("second tier" cities) and particular opportunities are thought to exist in the transport sector and from new sources of Big Data, including Google Maps (Jindal, Kumar, & Singh, 2020; Rakesh, Heeks, Chattapadhyay, & Foster, 2018; Rizwan, Suresh, & Rajasekhara Babu, 2016). Hubballi-Dharwad, the case study analysis focuses on among the 100 cities in the "Smart Cities Mission", and "Smart Mobility" is recognised as a key area for intervention (Hubballi-Dharwad, 2013).

A growing body of research has considered the role of Big Data for transport research and planning (Batty, 2013; Calabrese et al. 2013; Milne & Watling, 2019; Tzika-Kostopoulou & Nathanail, 2021) with a branch of this research focusing specifically on Google Maps data (Hanna, Kreindler, & Olken, 2017; Kreindler, 2016; Dumbliauskas, Grigonis, & Barauskas, 2017; Akbar & Duranton, 2017). This article adds to this research in two ways.

First, by focusing on a smaller urban centre in the Global South, this research considers an underexplored context. In contrast with the larger and often wealthier urban centres that are the focus of much existing research, smaller urban centres frequently have very high urban growth rates and are yet to invest significantly in public transport networks. Such urban centres are also more likely to face capacity issues in government due to smaller budgets and less established institutional structures, possibly leading them to be more attracted to "smart innovations" using Big Data. Cities with these characteristics are anticipated to be the source of the majority of urban population growth over the coming decades (UN DESA 2019) and are, therefore, critical to the achievement of the Sustainable Development Goals. Sustainable urban mobility plans (SUMPs) are a focus of a growing body of literature (Okraszewska et al., 2018); however, low-income regions of the world continue to be underrepresented.

The potential value of Big Data for transport planning in Hubballi-Dharwad, and "second-tier" global cities more generally, is considered in the first and second sections of the Results. We first analyse the extent to which big data derived from Google Maps can provide information on key attributes of the transport system, including hourly and weekly travel times on key routes, before assessing the impact of a new bus-rapid transport line on travel times to key locations in the city.

The second way this analysis adds to the existing literature is by probing some of the specific potential shortcomings of Big Data raised by existing authorship. The third section of the Results assesses the possibility that algorithms may capture, sort, clean, and pass on data in ways that alter our understanding of phenomena (Zou & Schiebinger, 2018) by focusing on a major rainstorm event that affected Hubballi-Dharwad in June 2018. Finally, the fourth section of the Results assesses potential biases in the data (Kwan, 2018) by comparing the quality of the data provided from informal settlement and wealthier parts of the city.

2 | STUDY AREA

A rapidly expanding urban population and sprawling cities are placing increasing pressure on transport systems in India. At the same time, partly as a result of increasing incomes, there is a growing trend towards private transport. The transport sector contributes to about 15% of CO_2 emissions in India, a share that has been increasing over time (Gupta & Garg, 2020) and congestion, air pollution, and road

traffic accidents are common in urban areas, at great cost to society and the economy (Rajasekaran, Rajasekaran, & Vaishya, 2021).

The National Urban Transport Policy (NUTP) of 2006 emphasised the need to give greater priority to public transport, and the Sustainable Urban Transport Programme (SUTP) was designed to support and demonstrate the principles of the NUTP. Following the adoption of these policies, a Bus Rapid Transit (BRT) scheme connecting the twin cities of Hubballi and Dharwad was chosen as a demonstration project. The engineering study completed before construction contains many of the "best practice" elements for BRT networks. For example, the system is designed with a dedicated roadway, raised platforms, a limited number of stations, and an electronic payment system. Importantly, the document also highlights that reducing congestion along the main corridors of the city is a key justification for the project (CEPT, 2013). Assessing private vehicle travel times along the route is, therefore, seen as an indirect means of assessing the success of the project and its overall impact on the city's transport network.

Whether Google Maps travel time estimates (or other Big Data sources) can be used in this way has relevance beyond Hubballi-Dharwad. BRT systems are considered an important tool for climate change mitigation due to their potential to provide an important public service while also contributing to global emissions reduction targets (Gouldson et al., 2020; Sudmant, Mi, et al., 2020; Sudmant, Verlinghieri, et al., 2020). Studies from several cities with well-established BRT systems - such as Bogota, Johannesburg, and Mexico City (Ingvardson & Nielsen, 2018), substantiate this. In addition, BRTs contribute to the reduction of air pollutants such as carbon monoxide and particulate matter, primarily through reducing the total number of vehicle kilometres travelled and by encouraging the replacement of older, smaller vehicles with newer, cleaner high-capacity buses (Stankov et al., 2020). Research also suggests that BRT systems can contribute to equity objectives by providing low-income groups with greater access to public transport, travel time and cost savings, and safety benefits (Venter, Jennings, Hidalgo, & Pineda, 2018).

With a population of 940,000, Hubballi-Dharwad municipal area is the second largest urban agglomeration in Karnataka State after Bangalore (Figure 1). Hubballi is the region's commercial centre while Dharwad is the administrative and educational hub (UNESCAP, 2014). A BRT was proposed as a way to improve connectivity between the cities - which are around 20 km apart - and to temper vehicular growth, while accommodating an urban population that is projected to reach almost 1.5 million by 2030 (ibid). After delays resulting from complex land acquisition processes, the Hubballi-Dharwad BRT began operations in October 2018.

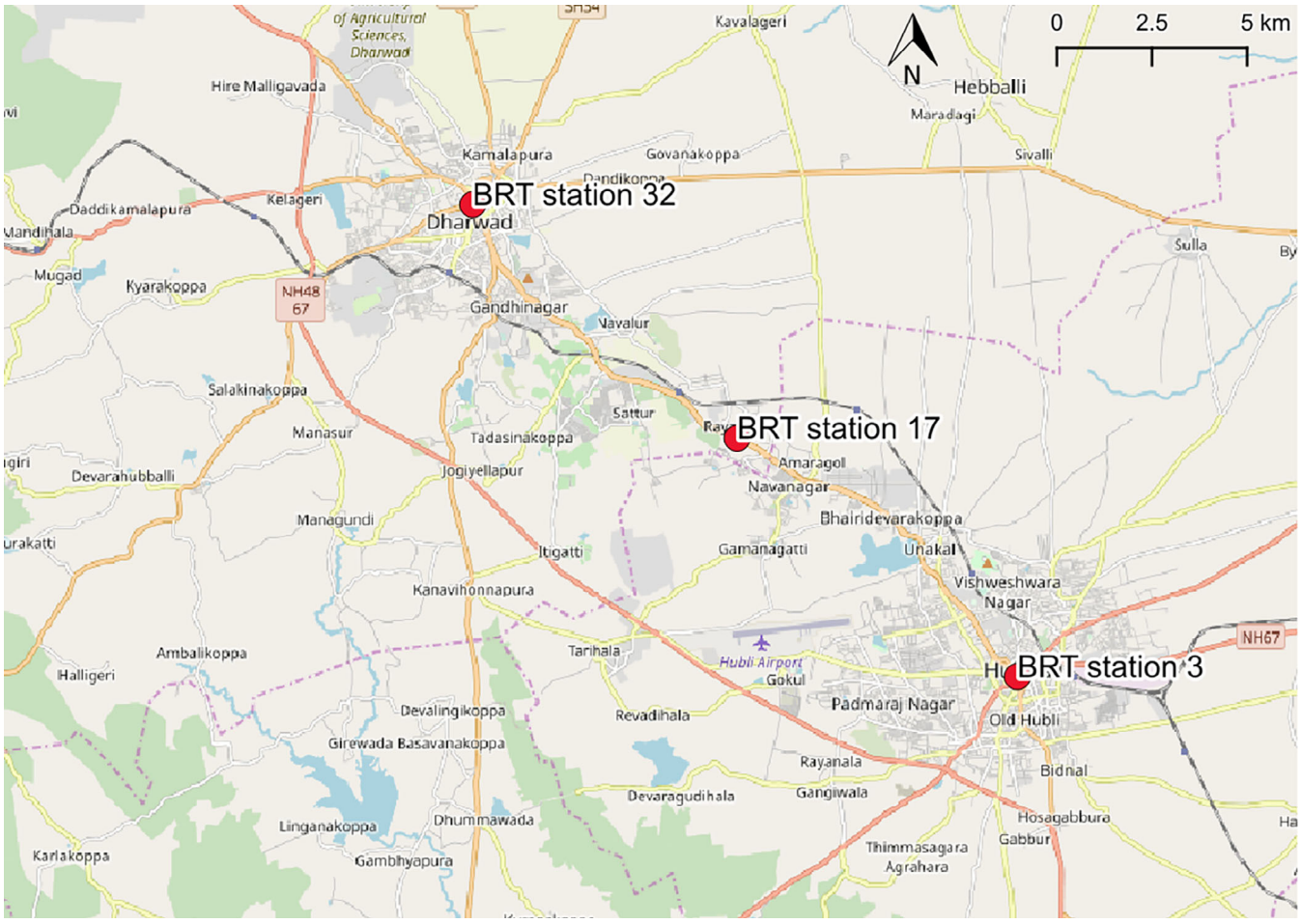

FIGURE 1 Map of Hubballi-Dharwad and the BRT stations studied [Colour figure can be viewed at wileyonlinelibrary.com]

Almost one-fifth of the population of Hubballi-Dharwad lives in informal settlements (Ministry of Housing and Urban Affairs, 2019). This is significant in the context of this research since informal settlements have fewer vehicles that Google can track to determine travel times, and residents may have fewer devices from which Google can collect data. A recent study from Karnataka's capital, Bangalore, shows that less than 1% of informal settlement households own a car (Roy et al., 2018), compared with more than 70% of wealthier households in the same city (Bansal, Kockelman, Schievelbein, & Schauer-West, 2018). Assuming this is representative of cities in India, this could make it potentially challenging for Google to estimate travel times from these areas, a matter we investigate at the end of the Results section.

3 | METHODOLOGY

A set of measurements was conducted through Google Maps Application Programming Interface (API) queries with information collected on estimated live travel trip durations and distances. Three of the 33 new bus stations that are part of the BRT were selected as departure and arrival points: The Chennamma station located in the centre of Hubballi; ISKCON located in the middle of the BRT pathway and near the main hospital; and Jubilee Circle in the centre of Dharwad. In addition, an analysis grid of 1 and 2 km resolutions covering the urban area of the city was built around the bus stations comprising 206 grid cells (Figure 2). Queries for both the trip time and distance from each grid cell to each of the centres and each centre to each grid cell were conducted every hour from 5 a.m. to 11 p.m. from April 2018 to mid-July 2018. From mid-July 2018, queries were limited to Tuesday and Thursday until mid-February 2019.

Although Google Maps' estimated time of arrival algorithm is not public, it is understood that Google uses different features to assess live travel times. These include official speed limits, recommended speeds, information on road types, and topography and real-time traffic information. A mix from these different data is processed to enhance the algorithm.

4 | RESULTS

Results are presented in four sections. First, Google Maps travel times estimate data are used to identify key attributes of the transport network, including hourly and weekly traffic variation across the city.

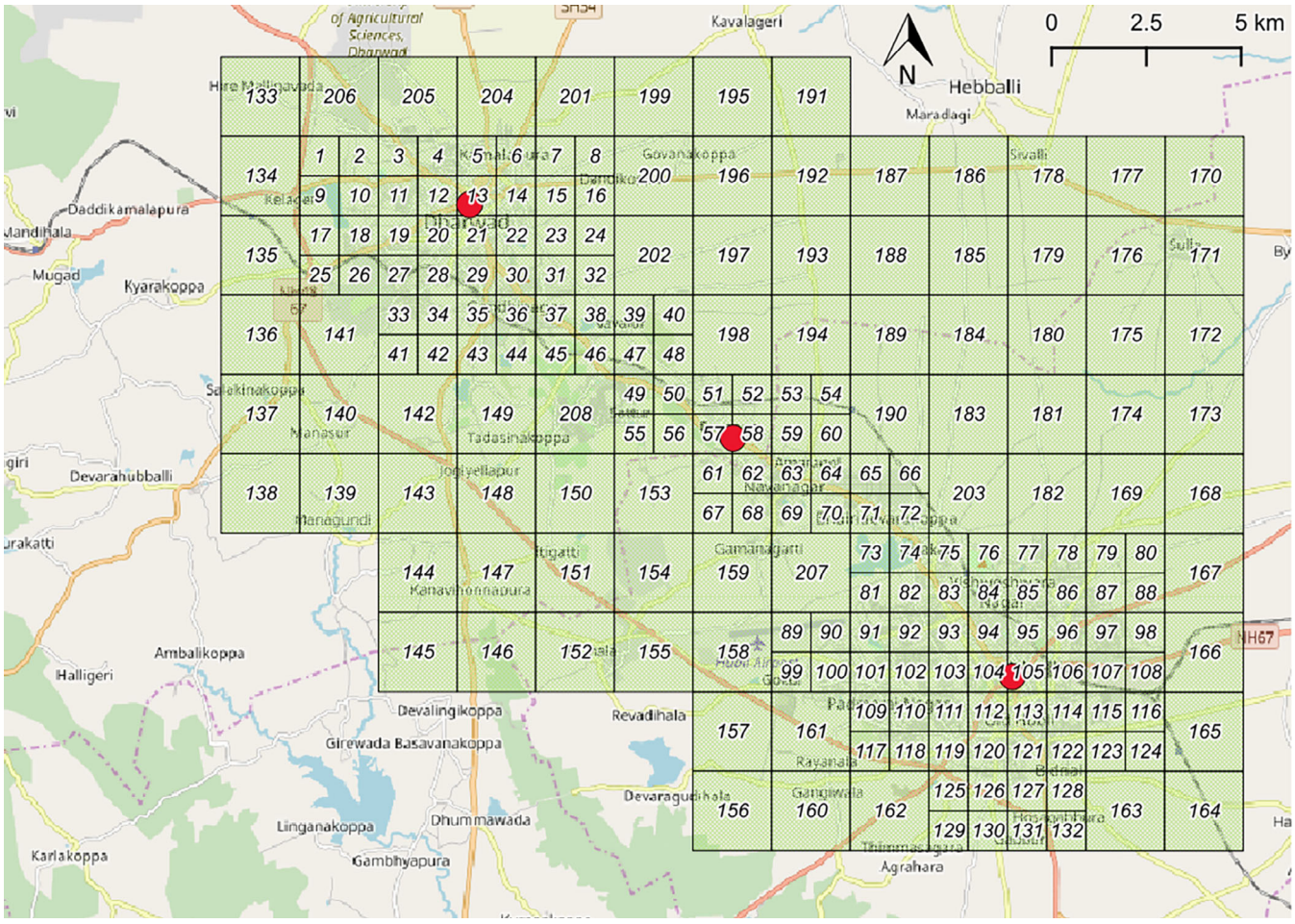

FIGURE 2 Grid points used in the analysis [Colour figure can be viewed at wileyonlinelibrary.com]

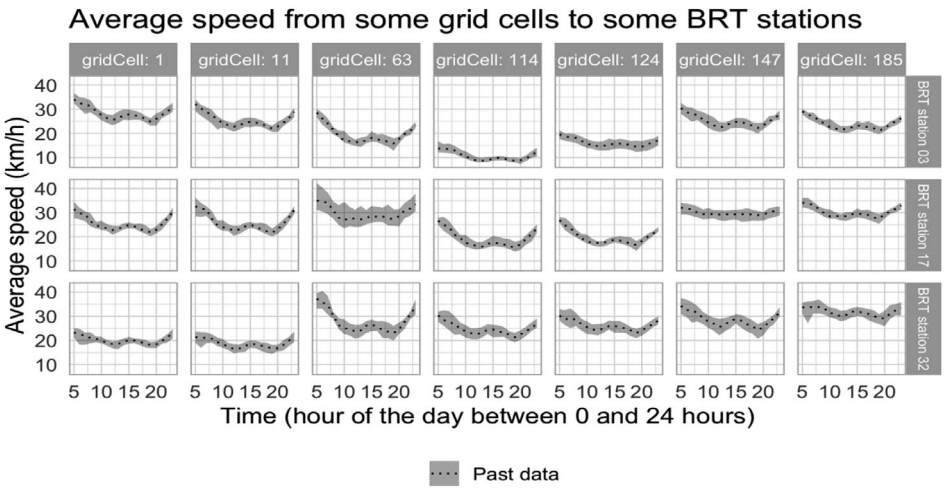

FIGURE 3 Average speed by time of day for different routes and grid cells and the 95% confidence interval (grey)

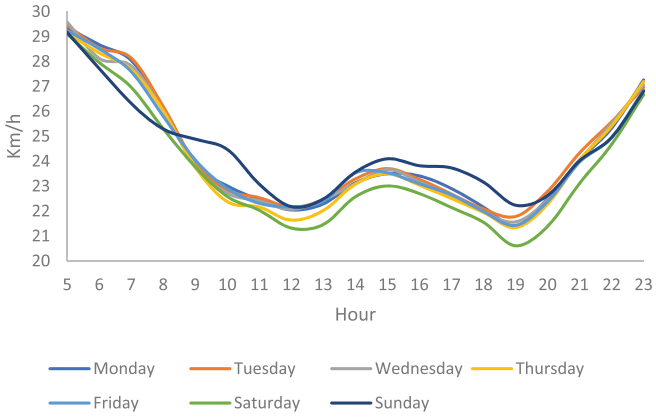

FIGURE 4 Average speed by time of day and day of week [Colour figure can be viewed at wileyonlinelibrary.com]

Second, we investigate the effect of the new BRT project on vehicle travel times. Third, we assess how Google Maps capture the impact of monsoon flooding on June 4th, 2018. Fourth, we compare Google travel estimate data with a simple model of travel times in the city to assess the extent that Google travel data are providing "additional value" beyond more basic modelling approaches, and the extent that biases may be present in the data.

4.1 | Traffic variation in Hubballi-Dharwad

Figure 3 shows average travel speeds between grid cells and BRT stations estimated using the travel times estimates from Google and the trip distance provided by Google. Results show a similar pattern over time, with congestion slowing travel speeds between 9 a.m. and 2 p.m. (approximately) and then again between 5 p.m. and 9 p.m. (approximately). However, the degree of congestion (the change in speed) and the speed under low congestion are found to depend significantly on factors specific to different routes.

In Figure 4, the morning and evening congestion periods are presented more clearly by assessing travel speeds across different routes. Combining these by day of the week reveals that Saturday has the most traffic and Sunday has the least traffic congestion. The effect of time of day is seen to be significantly more important than the day of the week for the level of congestion given the much larger differences in travel speeds.

4.2 | The impact of the bus-rapid transport network

In order to understand the effect of the BRT on travel times, we assess travel times before and after the BRT began operation and compare routes that are parallel to and perpendicular to the BRT. The hypothesis behind this approach is that trips parallel to (or along) the BRT line will be affected by the new transport option, while trips perpendicular to the BRT should not be affected. To provide clarity, the city is divided into regions, as shown in Figure 5.

Figures 6–8 show the change in travel times before (blue dots) and after (red dots) the BRT. Results show a statistically significant (at the 5% level), but small, change in travel times for most routes parallel to the BRT. Note, those trips parallel to the BRT include all trips represented on the top row of each figure, labelled along, as well as some of the routes in the second row. Routes not parallel to the BRT, by contrast, do not show a consistent change in travel times. These findings suggest that in the immediate weeks and months after the implementation of the BRT, the new bus has had the effect of improving congestion, one of the stated goals of the project. However, whether this effect is by moving drivers from cars onto the bus, by discouraging drivers from taking this route, or by another means is beyond the scope of this analysis to determine. Further, who is taking the BRT and how the specific trips they are taking have been affected, is information not available using this data set and approach.

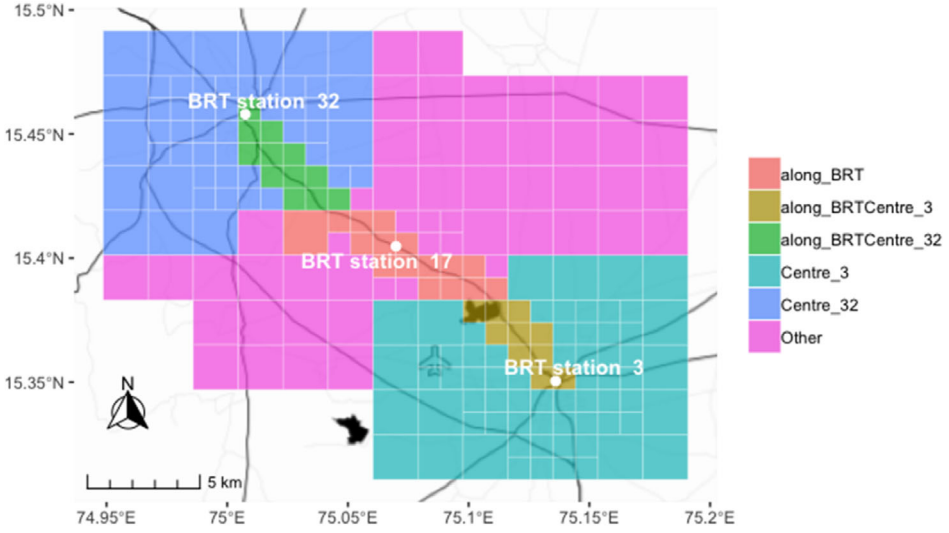

FIGURE 5 Zones of the city used to establish the impact of the BRT [Colour figure can be viewed at wileyonlinelibrary.com]

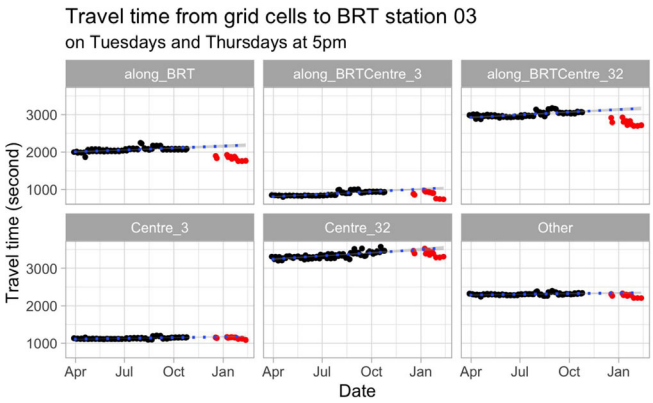

FIGURE 6 Travel times from grid cells to BRT station 03. Please note, the route along the new BRT offers to most direct route from "Centre 32" to station 3 [Colour figure can be viewed at wileyonlinelibrary.com]

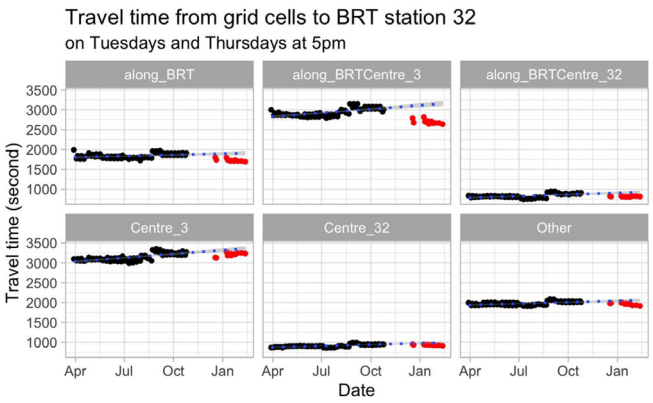

FIGURE 8 Travel time from grid cells to BRT station 32. Please note, along the BRT route offers the most direct route from Centre 3 to BRT station 32 [Colour figure can be viewed at wileyonlinelibrary.com]

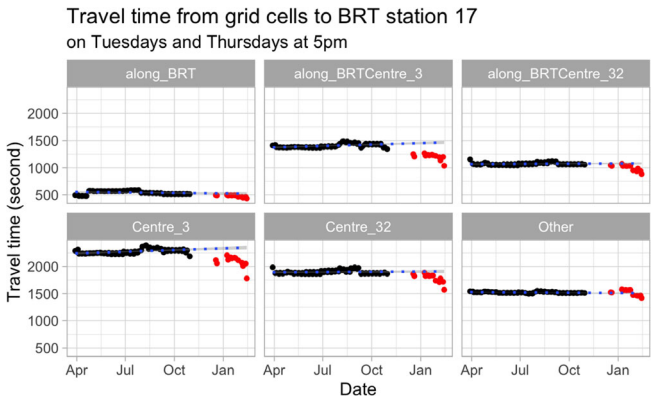

FIGURE 7 Travel times from grid cells to BRT station 17. Please note, along the BRT route offers to most direct route from both 'Centre 3' and 'Centre 32' to BRT station 17 [Colour figure can be viewed at wileyonlinelibrary.com]

4.3 | "Real-time" traffic analysis

One of the key advantages of having access to Big Data is its ability to rapidly provide information for users. To assess whether Google Maps transport data could inform transport policy-making in acute situations, we look at data from June 4th, 2018, a day of heavy rain and flooding in Hubballi-Dharwad that followed on several previous days of heavy rain and flooding in the region (The Times of India, 2019).

Figure 9 shows combinations of travel times on Monday, June 4th, 2018 on different routes at different times of day, and the average speed for that route at that time of day on Mondays (excluding June 4th, 2018), across the dataset. Despite major flooding, results suggest that travel speeds across the city on Monday, June 4th, 2018 were very similar to travel speeds on a typical Monday and not significantly different at the 5% level. Similarly, there is little evidence of disruption to any specific routes. Of the 25 observations that showed

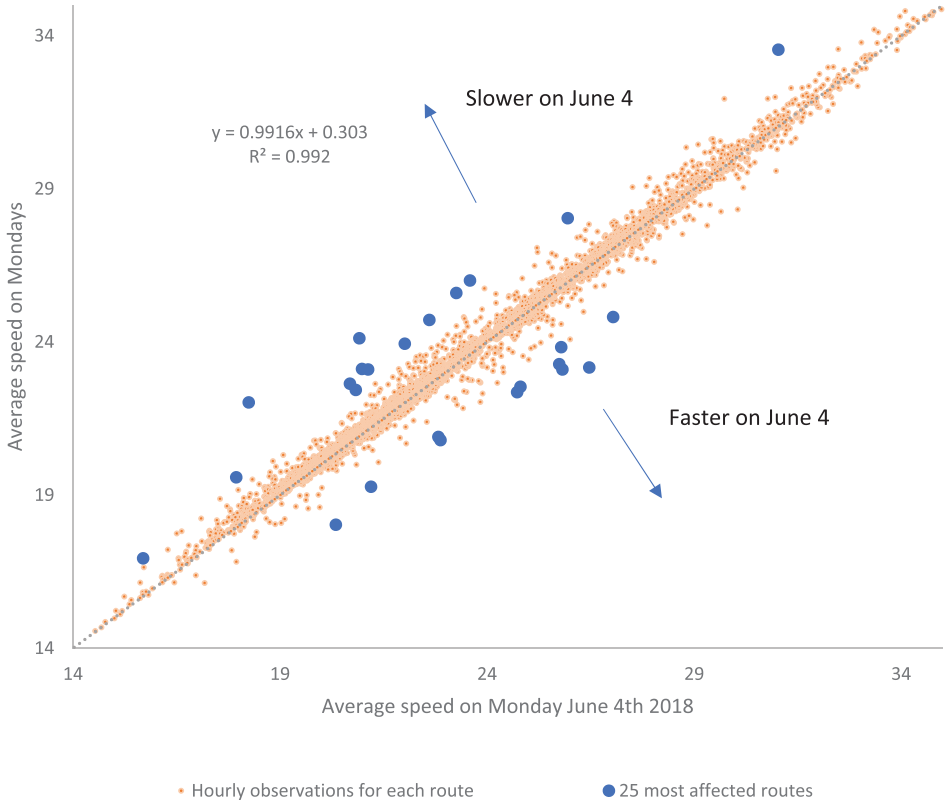

FIGURE 9 Speed for each hour-route combination on an average Monday and on Monday, June 4th [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Adjusted R^2 for models applied to data from grid cells (regions of the city) with the highest proportion of informal settlement dwellers and from the remainder for the city

		Trip distance (km)			
	Model	<2 ($n = 86{,}559$)	<5 ($n = 505{,}702$)	<10 ($n_o = 1{,}149{,}861$)	<15 ($n = 1{,}956{,}916$)
90% of grid cells with the lowest proportion of informal settlement dwellers	Distance	32%	60%	68%	78%
	Distance and density	43%	64%	69%	78%
	Distance and density and hour dummies	53%	71%	74%	83%
	Distance and density and hour dummies and day dummies	53%	71%	75%	83%
	Distance and density and hour dummies and day dummies and restricted to 1 km by 1 km cells	53%	72%	75%	87%
	Model	**<2 ($n = 2{,}810$)**	**<5 ($n = 35{,}505$)**	**<10 ($n = 95{,}023$)**	**<15 ($n = 170{,}615$)**
10% of grid cells with the highest proportion of informal settlement dwellers	Distance	26%	69%	66%	75%
	Distance and density	26%	72%	66%	76%
	Distance and density and hour dummies	85%	87%	74%	82%
	Distance and density and hour dummies and day dummies	87%	87%	75%	82%
	Distance and density and hour dummies and day dummies and restricted to 1 km by 1 km cells	87%	87%	79%	89%

Note: Number of observations does not apply for the model specification that includes only 1 km by 1 km cells.

the largest impact (each observation is one data point, representing an estimation of the travel time and distance between a cell and a centre or a centre and a cell), 14 showed faster times on June 4th and 9 showed slower times. Of these, only five routes were 10% faster or slower than usual.

4.4 | Comparing Google maps estimates with a simple transport model

Following the results in the previous section, we were curious to explore the extent to which Google data is adding on-the-ground information to its estimates. In the absence of detailed information on the way Google Maps estimates are calculated, we develop a model of travel times that is based on a set of characteristics seen to have an important role in predicting travel times: the hour of the day, the day of the week, population density, and the distance of the trip (Table 1). Using linear-regression, this model explains 85% of the variation across all 3.2 million trips in our dataset, suggesting that 15% of the variation in Google's estimated travel times is related to other factors. We assume that a significant portion of this 15% of addition variation comes from Google's ability to collect real-time travel information on actual travel conditions, relating, for example, to the weather, traffic accidents, or other events that are too rare or uncertain to be included in the characteristics of our model.

The extent to which Google is able to capture this real-time information may not be the same across the city, particularly in informal settlement areas due to a lower concentration of mobile devices. To test this hypothesis, we can compare the fit of our model for trips starting from informal settlement areas versus the fit of our model for a trip starting from non-informal settlement areas.

If a subset of the dataset (informal settlement or non-informal settlement originating trips) shows a lower R^2 in our model, this suggests that Google might have more real-time information, allowing Google to provide more bespoke travel time estimates that differ from the ones in the "basic model." If the R^2 is higher, this suggests Google travel time estimates are more likely to be based on a set of characteristics similar to those in our model, implying that they may not have more information to improve their estimates. This effect should be magnified for shorter trips. Longer trips will frequently converge onto the same routes and over the course of a longer trip, drivers will have more opportunity to change their route to avoid traffic. We would therefore expect the R^2 to be higher for relatively long trips compared with shorter trips.

Results find that the model of travel times we apply explains a higher proportion of all variation in trip times from informal settlement areas compared with the remaining grid cells. This phenomenon exists across all grid cells and also when we restrict our analysis to the "finer" 1 km square cells. Results also show a higher R^2 as the minimum trip length is increased, in line with our assumption about longer trips.

These findings could be a result of fundamental aspects of transport in Hubballi-Dharwad. Travel times from informal settlement areas may be more predictable due to geography or the configuration of the travel network. This would be despite the fact that informal settlement areas are found across Hubballi-Dharwad, including adjacent to formal settlement areas. However, without detailed information on the raw data Google Maps is using, or the way that data is processed before it is passed through Google Maps, we cannot rule out that the data we are being provided with is more detailed outside of informal settlement areas.

5 | DISCUSSION

From the perspective of a policymaker in Hubballi-Dharwad (or another developed or developing urban area), the analysis presented demonstrates what appear to be some clear benefits of using Google Maps data to inform the evaluation of transport policies, plans, and programmes. Compared with surveys, traffic counts, and other traditional methods of data collection, Google Maps makes it relatively easy to collect large quantities of data in a timely fashion across the entire timeline of a project, irrespective of weather, holidays, or other challenges.

Moreover, this data can be used in ways that have clear policymaking value. Information on travel times by time of day and day of week can inform public transport scheduling, road maintenance and public works, and long-term urban development planning. This kind of research is foundational for urban transport policymaking and planning, but the challenge of collecting data and building bespoke models is a barrier in high and low-income contexts alike.

Analysis of the BRT suggests Google Maps data may also be able support ex-post assessment, a process that is critical for learning but often not undertaken due to the cost and challenge of accessing data (Nicolaisen & Driscoll, 2014). Results here, which show a relatively modest change in travel times along the BRT compared with routes perpendicular to the BRT, also highlight the value of the large datasets accessible with Google Maps, which allow for a level of statistical robustness that would be challenging with other methods.

A transport department that completed these analyses could easily replicate them in the future. And since policymakers in other urban areas also using Google Maps would have access to data of the same types and format, knowledge sharing, and learning could be radically increased. These realisations have enchanted academics who forecast the beginning of a fundamental shift in our epistemological approach to transport planning led by data analysis rather than the development of hypotheses (Kitchin, 2014; Rabari & Storper, 2015) and suggest the private sector could play an important role supporting sustainable low carbon development (Colenbrander, Sudmant, Chilundika, & Gouldson, 2019; Scheyvens, Banks, & Hughes, 2016; Sudmant, Colenbrander, Gouldson, & Chilundika, 2017).

The extent that such a shift in the nature of urban transport and urban transport policymaking is on the horizon is beyond the scope of this paper. However, the third and fourth analyses in the results section were undertaken with the intention of exploring how Google data might contribute to more novel analysis of the kind that has been associated with this transition in transport planning (cf. Kitchin, 2014).

The speed with which data can be collected and assessed is a key feature of Big Data and has clear value for transport policymakers. Rapid analysis can help in identifying transport hotspots and responding to emergencies. In contrast with our personal experience with Google Maps in other urban contexts during periods of disruption, however, we were surprised to find no clear impact of the flooding in Hubballi-Dharwad and the surrounding region on Monday, June 4th, 2018 in the data.[1] This finding is not only suspicious but also alarming. Google Maps is widely used by residents in the area and could, therefore, have encouraged potentially dangerous travel decisions. It also raises concerns about the veracity of the earlier findings of this analysis: data varies by time of day, day of week, and distance, but to what extent is the data based on on-the-ground information?

The data provided by Google comes without any information on how it was put together. However, based on the limited information available about Google Maps' algorithm, we can assume that travel time estimates are derived from both "real" data collected from travellers with Android and Google devices, and from a model of urban transport used to make estimates in the absence of information from connected devices. The factors in this model may include trip distances, topography, and the time of day, among other factors.

In order to probe the characteristics of this underlying algorithm, we developed a simple model of the transport network. Across the entire dataset, results show that characteristics, including time of day, day of the week, the distance of a trip, and the density of the urban area travelled through describe 85% of the variation in travel times. This suggests that either these variables, or factors correlated with them, are constituents of the model used by Google. This also suggests that 15% of the variation in estimated travel times may be attributable to other variables or information captured by Google connected devices. Wider factors might include topography, road quality, and speed limits, while information collected from connected devices might include traffic caused by a car breaking down, a slow driver, or weather.

In this context, we would assume that data captured by Google connected devices would override the estimates of the model. Described another way, if Google has information that a road is poor quality, on a steep hill, and that it is the busiest day of the week and time of the day (implying that a road is likely to be relatively slow for vehicles according to the model), but connected devices are reporting that vehicles are travelling quickly, we would assume that Google would eventually conclude that this is a relatively fast route for cars and provide estimates accordingly. Similarly, for the opposite case, data from connected devices should allow Google to correctly predict slower travel times on roads even if an ex-ante estimate suggested relatively fast travel speeds. Over a long period of time, during which many data points are collected, Google estimates should improve significantly by this means.

Importantly, the extent to which Google can account for certain unpredictable events (e.g., a car breaking down) will likely still depend on timely data from connected devices. All else constant, this factor will be most prominent for shorter trips where there are fewer opportunities for alternative routes to avoid such events. We would, therefore, expect that the difference between a basic model of the transport network and a more complicated (and, by assumption, more accurate) model, such as that used by Google, would be largest for shorter trips and smallest for longer trips.

The failure of the just mentioned hypothesis for trips starting from informal settlement areas, with the model we have developed providing a similar degree of accuracy for shorter and longer trips, maybe explained in three (non-exclusive) ways. First, as with any statistical analysis, there is the potential that these results are a statistical artefact. This is mitigated to some degree by the number of observations and by the different specifications of the model presented. Second, the elements of the basic model may be a better fit for trips from informal settlements over shorter distances. In other words, characteristics left out of our model, including topography, weather, and car accidents may only have a small effect on travel times from informal settlement areas. This seems unlikely as the informal settlements are in different parts of the city and adjacent in many cases to wealthier areas (see Figure 2). Further, one would not expect some of these factors (a car breaking down or an unexpected rainstorm) to be significantly correlated with the wealth of the neighbourhood car passing through.

Finally, Google travel times from the informal settlement areas of the city may not include the same amount and quality of on-the-ground data as they are able to access from wealthier areas, forcing Google to provide less accurate estimates. We would emphasise that these results call for further research to be verified. However, there is reason to think, the third of these explanations could be the cause of these results. Only approximately one-third of the population has a smartphone in India in 2019 (Statista, 2019). The vast majority of these devices are Android, but ownership is skewed towards the wealthier population (ibid). And among the poorer population, some share a device or leave it at home for safety purposes, further reducing their visibility in data collected. These factors suggest that there is a causal pathway that could lead to lower quality travel time estimates from poorer areas.

While concerns around systematic biases in Big Data sets are well established (Batty et al. 2012; Kwan, 2018), a number of authors have implicitly made the assumption that these biases are not large enough to be a concern in analyses of Google data. In addition, the exact nature of these biases remains poorly explored. Here, we find some evidence to suggest the existence of spatial and temporal limitations of Google data, which may have a social consequence: reduced quality of travel time data for informal settlement populations with implications for urban policymaking, and inclusive urban development.

It should be noted that on-the-ground assessment to confirm these findings, or comparison with a city-based transport model, was not possible. Nonetheless, these latter analyses raise wider concerns about the use of Big Data for informing urban policies, plans, and programmes. If there is no transparency around the quality of data and the way it has been processed there may be significant limits to the extent that surprising results can be explained, leading to concerns about datasets as a whole. This issue is particularly evident in our

findings around the days Hubballi-Dharwad faced flooding but apply also to the findings on the differences between informal and non-informal settlements, and the impact of the BRT. And since the data available is wide but thin, that is, massive in the quantity of information but lacking in number of variables, corroborating the results with other datasets is challenging.

Important in this context is that the potential for errors in the data is known, but the nature of these errors is not. This is in contrast with conventional transport modelling methods where the exact nature of errors is unknown, but comparisons with other datasets can be used to determine confidence levels and indications of bias. Big Data sources rarely come with a detailed methodology, quality assurance, or user manual of any kind. On the contrary, Big Data is often described as speaking for itself (Villanueva et al., 2016). But, if the data is of questionable validity—and therefore, does not speak for itself—there may be some irony in using it for ex-post analysis.

For policymakers, the key concern in this context regards unintended consequences. An individual's travel app that does not work during poor weather may lead to a dangerous travel decision, but more likely leads only to a lengthy commute. A transport planner basing a policy or project on data that only considers fair weather, by contrast, may lead to a city gridlocked for the course of the monsoon.

For the academic community, the specific aspects of urban life that are misrepresented or that fall between the columns of ever more impressive datasets may be a secondary, if critical, issue. Faced with a new age of seemingly limitless information, more fundamental questions may consider the ways algorithmic governance expands the capacity to govern by replacing or crowd out other forms of knowledge and power.

Reflecting on waves of enthusiasm for more "scientific" approaches to urban planning over recent decades, Duminy and Parnell (2020) remind us that the debates between "interpretivists" and "positivists" are old, well defined, and possibly growing more acrimonious. A practical path forward may lie with efforts to emphasise the value in different ways of generating urban knowledge. Big Data provides increasing rapid and analytically robust means of addressing specific questions. The robustness of the framing of those questions, and whether results hold wider significance, however, may be better informed by a wider plurality of urban methods and approaches, including the citizen science movements (Callaghan, Poore, Major, Rowley, & Cornwell, 2019), citizens' assemblies (Van Crombrugge, 2020), urban labs (Acuto, Dickey, Butcher, & Washbourne, 2021), participatory games (Andreotti, Speelman, Van den Meersche, & Allinne, 2020), and a burgeoning set of wider methods that are being applied in a growing number of urban areas (Creasy, Lane, Owen, Howarth, & van der Horst, 2021).

6 | CONCLUSIONS

Google Maps and other sources of Big Data present an emerging opportunity for policymaking in transport and more widely. The extent to which these approaches can be relied upon, however, depends on the value they add to analysis weighed against the new limitations and sources of uncertainty they generate. To date, quantitative analysis has placed a much greater focus on the opportunities. Here, we contribute to what we hope will be a growing field of analysis assessing the quantitative shortcomings of Big Data approaches for informing policymaking, and how these may be overcome, where efforts are made to understand the lived realities behind the data and the complementarities between Big Data and wider methods of knowledge generation in urban areas.

Future analysis in this field can be targeted to three areas. First, analysis can explore the existence and extent of disparities between the value of Big Data for populations from different socio-economic backgrounds. This analysis is essential to understand the extent and possible consequences for sustainable development, especially in rapidly growing urban areas where the vast majority of infrastructure is yet to be built. Second, analysis is needed to "truth" the proliferation of Big Data sources with on-the-ground realities. This can help to determine the key areas new data sources have shortcomings and advantages relative to established sources of information and methods of analysis. Finally, interdisciplinary work that explores, both conceptually and in practical terms, the ways empirical and qualitative urban data sources can be integrated is needed to ensure wider methods of knowledge generation in urban areas can complement the growing proliferation of Big Data.

ACKNOWLEDGEMENT

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article. This work was supported by funding from the Department for International Development grant 113550.

ORCID

Andrew Sudmant https://orcid.org/0000-0001-8650-8419
Andy Gouldson https://orcid.org/0000-0002-1464-6465

ENDNOTE

[1] During the preceding 3 days and continuing through Monday night the wider area and Hubballi-Dharwad received several meters of rain. Flooding during this period made it one of the most severe monsoon seasons on record (The Times of India, 2019).

REFERENCES

Acuto, M., Dickey, A., Butcher, S., & Washbourne, C.-L. (2021). Mobilising urban knowledge in an infodemic: Urban observatories, sustainable development and the COVID-19 crisis. *World Development*, *140*, 105295. https://doi.org/10.1016/j.worlddev.2020.105295

Akbar, P., & Duranton, G. (2017). *Measuring the cost of congestion in highly Congested City: Bogotá* (Working Paper). CAF. Retrieved from http://scioteca.caf.com/handle/123456789/1028

Albino, V., Berardi, U., & Dangelico, R. M. (2015). Smart cities: Definitions, dimensions, performance, and initiatives. *Journal of Urban Technology*, *22*(1), 3–21. https://doi.org/10.1080/10630732.2014.942092

Andreotti, F., Speelman, E. N., Van den Meersche, K., & Allinne, C. (2020). Combining participatory games and Backcasting to support collective scenario evaluation: An action research approach for sustainable

agroforestry landscape management. *Sustainability Science*, *15*(5), 1383–1399. https://doi.org/10.1007/s11625-020-00829-3

Bansal, P., Kockelman, K. M., Schievelbein, W., & Schauer-West, S. (2018). Indian vehicle ownership and travel behavior: A case study of Bengaluru, Delhi and Kolkata. *Research in Transportation Economics*, *71*, 2–8. https://doi.org/10.1016/j.retrec.2018.07.025

Batty, M. (2012). Smart cities, big data. *Environ Plann B Plann Des 39*, 191–193. https://doi.org/10.1068/b3902ed.

Batty, M. (2013). Big data, smart cities and city planning. *Dialogues in Human Geography*. https://doi.org/10.1177/2043820613513390.

Behbahani, H., Nazari, S., Kang, M. J., & Litman, T. (2019). A conceptual framework to formulate Transportation network design problem considering social equity criteria. *Transportation Research Part A: Policy and Practice*, *125*, 171–183. https://doi.org/10.1016/j.tra.2018.04.005

Bissell, D. (2018). Automation interrupted: How autonomous vehicle accidents transform the material politics of automation. *Political Geography*, *65*, 57–66. https://doi.org/10.1016/j.polgeo.2018.05.003

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., & Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation Research Part C: Emerging Technologies, *26*, 301–313. https://doi.org/10.1016/j.trc.2012.09.009.

Cabannes, Y., & Lipietz, B. (2018). *The democratic contribution of participatory budgeting* (Working paper no. 34).

Callaghan, C. T., Poore, A. G. B., Major, R. E., Rowley, J. J. L., & Cornwell, W. K. (2019). Optimizing future biodiversity sampling by citizen scientists. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1912), 20191487. https://doi.org/10.1098/rspb.2019.1487

CEPT. (2013, March). *Bus rapid transit system Huballi-Dharwad detailed feasibility report annexure*. CEPT University, Ahmedabad.

Colenbrander, S., Sudmant, A., Chilundika, N., & Gouldson, A. (2019). The scope for low-carbon development in Kigali, Rwanda: An economic appraisal. *Sustainable Development*, *27*, 349–365.

Colenbrander, S., Sudmant, A. H., Gouldson, A., de Albuquerque, I. R., McAnulla, F., & de Sousa, Y. O. (2017). The economics of climate mitigation: Exploring the relative significance of the incentives for and barriers to low-carbon Investment in Urban Areas. *Urbanisation*, *2*(1), 38–58.

Coletta, C., & Kitchin, R. (2017). Algorhythmic governance: Regulating the 'heartbeat' of a city using the internet of things. *Big Data & Society*, *4*(2), 2053951717742418. https://doi.org/10.1177/2053951717742418

Cookson, G., & Pishue, B. (2017). *INRIX global traffic scorecard*–Appendices. p. 38.

Creasy, A., Lane, M., Owen, A., Howarth, C., & van der Horst, D. (2021). Representing 'place': City climate commissions and the institutionalisation of experimental governance in Edinburgh. *Politics and Governance*, *9*(2), 64–75.

Creutzig, F., & He, D. (2009). Climate change mitigation and co-benefits of feasible transport demand policies in Beijing. *Transportation Research Part D: Transport and Environment*, *14*(2), 120–131. https://doi.org/10.1016/j.trd.2008.11.007

Docherty, I., Marsden, G., & Anable, J. (2018). The governance of smart mobility. *Transportation Research Part A: Policy and Practice*, *115*, 114–125. https://doi.org/10.1016/j.tra.2017.09.012

Driscoll, P. A. (2014). Breaking carbon lock-in: Path dependencies in large-scale transportation infrastructure projects. *Planning Practice & Research*, *29*(3), 317–330. https://doi.org/10.1080/02697459.2014.929847

Dumbliauskas, V., Grigonis, V., & Barauskas, A. (2017). Application of google-based data for travel time analysis: Kaunas City case study. *Promet-Traffic &Transportation*, *29*(6), 613–621. https://doi.org/10.7307/ptt.v29i6.2369

Duminy, J., & Parnell, S. (2020). City science: A chaotic concept–And an enduring imperative. *Planning Theory & Practice*, *21*(4), 648–655. https://doi.org/10.1080/14649357.2020.1802155

Dwevedi, R., Krishna, V., & Kumar, A. (2018). Environment and big data: Role in smart cities of India. *Resources*, *7*(4), 64. https://doi.org/10.3390/resources7040064

Flyvbjerg, B., Skamris Holm, M. K., & Buhl, S. L. (2003). How common and how large are cost overruns in transport infrastructure projects? *Transport Reviews*, *23*(1), 71–88. https://doi.org/10.1080/01441640309904

Gouldson, A., Sudmant, A., Khreis, H., & Papargyropoulou, E. (2020). The economic and social benefits of low-carbon cities: A systematic review of the evidence (Working paper no. 92).

Gupta, D., & Garg, A. (2020). Sustainable development and carbon neutrality: Integrated assessment of transport transitions in India. *Transportation Research Part D: Transport and Environment*, *85*, 102474. https://doi.org/10.1016/j.trd.2020.102474

Hanna, R., Kreindler, G., & Olken, B. A. (2017). Citywide effects of high-occupancy vehicle restrictions: Evidence from 'three-in-one' in Jakarta. *Science*, *357*(6346), 89–93. https://doi.org/10.1126/science.aan2747

Hu, W., & Jin, P. J. (2017). An adaptive Hawkes process formulation for estimating time-of-day zonal trip arrivals with location-based social networking check-in data. *Transportation Research Part C: Emerging Technologies*, *79*, 136–155. https://doi.org/10.1016/j.trc.2017.02.002

Huballi-Dharwad. (2013). Retrieved 10 July, 2019 from, https://www.hubballidharwadsmartcity.com/smart-city-projects.html>

Hughes, S., Giest, S., & Tozer, L. (2020). Accountability and data-driven urban climate governance. *Nature Climate Change*, *10*(12), 1085–1090. https://doi.org/10.1038/s41558-020-00953-z

Ingvardson, J. B., & Nielsen, O. A. (2018). Effects of new bus and rail rapid transit systems–An international review. *Transport Reviews*, *38*(1), 96–116. https://doi.org/10.1080/01441647.2017.1301594

Jindal, A., Kumar, N., & Singh, M. (2020). A unified framework for big data acquisition, storage, and analytics for demand response management in smart cities. *Future Generation Computer Systems*, *108*(July), 921–934. https://doi.org/10.1016/j.future.2018.02.039

Khan, H. H., Malik, M. N., Zafar, R., Goni, F. A., Chofreh, A. G., Klemeš, J. J., & Alotaibi, Y. (2020). Challenges for sustainable Smart City development: A conceptual framework. *Sustainable Development*, *28*(5), 1507–1518. https://doi.org/10.1002/sd.2090

Kitchin, R. (2014). The real-Time City? Big data and smart urbanism. *GeoJournal*, *79*(1), 1–14. https://doi.org/10.1007/s10708-013-9516-8

Kitchin, R., Lauriault, T. P., & McArdle, G. (2015). Knowing and governing cities through urban indicators, City benchmarking and real-time dashboards. *Regional Studies, Regional Science*, *2*(1), 6–28. https://doi.org/10.1080/21681376.2014.983149

Kreindler, G. (2016). Driving Delhi? Behavioural responses to driving restrictions. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2966797

Kwan, M.-P. (2018). Algorithmic geographies: Big data, algorithmic uncertainty, and the production of geographic knowledge. *Annals of the American Association of Geographers*, *106*(2), 274–282. https://doi.org/10.4324/9781315266336-4

Litman, T. (2018). Toward more comprehensive evaluation of traffic risks and safety strategies. *Research in Transportation Business & Management*, *29*, 127–135. https://doi.org/10.1016/j.rtbm.2019.01.003

Mao, X., Yang, S., Liu, Q., Jianjun, T., & Jaccard, M. (2012). Achieving CO2 emission reduction and the CO-benefits of local air pollution abatement in the Transportation sector of China. *Environmental Science & Policy*, *21*, 1–13. https://doi.org/10.1016/j.envsci.2012.03.010

Milne, D., & Watling, D. (2019). Big data and understanding change in the context of planning transport systems. *Journal of Transport Geography*, *76*, 235–244. https://doi.org/10.1016/j.jtrangeo.2017.11.004.

Ministry of Housing and Urban Affairs. (2019). *Housing and slums: Hubballi-Dharwad*. Retrieved 08 August, 2019 from https://smartcities.data.gov.in/catalog/housing-and-slums-hubballi-dharwad?

filters%5Bfield_catalog_reference%5D=2911596&format=json&offset=0&limit=9&sort%5Bcreated%5D=desc

Nicolaisen, M. S., & Driscoll, P. A. (2014). Ex-post evaluations of demand forecast accuracy: A literature review. *Transport Reviews*, *34*(4), 540–557. https://doi.org/10.1080/01441647.2014.926428

Okraszewska, R., Romanowska, A., Wołek, M., Oskarbski, J., Birr, K., & Jamroz, K. (2018). Integration of a multilevel transport system model into sustainable urban mobility planning. *Sustainability*, *10*(2), 479. https://doi.org/10.3390/su10020479

Rabari, C., & Storper, M. (2015). The digital skin of cities: Urban theory and research in the age of the sensored and Metered City, ubiquitous computing and big data. *Cambridge Journal of Regions, Economy and Society*, *8*(1), 27–42. https://doi.org/10.1093/cjres/rsu021

Rajasekaran, R. B., Rajasekaran, S., & Vaishya, R. (2021). The role of social advocacy in reducing road traffic accidents in India. *Journal of Clinical Orthopaedics and Trauma*, *12*(1), 2–3. https://doi.org/10.1016/j.jcot.2020.12.021

Rakesh, V., Heeks, R., Chattapadhyay, S., & Foster, C. (2018). *Big data and urban transportation in India: A Bengaluru Bus Corporation Case Study*. https://doi.org/10.13140/RG.2.2.36761.77928

Rizwan, P., Suresh, K., & Rajasekhara Babu, M. (2016). *Real-time smart traffic management system for smart cities by using internet of things and big data*. 2016 International Conference on Emerging Technological Trends (ICETT), pp. 1–7. https://doi.org/10.1109/ICETT.2016.7873660

Roy, D., Palavalli, B., Menon, N., King, R., Pfeffer, K., Lees, M., & Sloot, P. M. A. (2018). Survey-based socio-economic data from slums in Bangalore, India. *Scientific Data*, *5*(1), 170200. https://doi.org/10.1038/sdata.2017.200

Scheyvens, R., Banks, G., & Hughes, E. (2016). The private sector and the SDGs: The need to move beyond 'business as usual': The private sector and the SDGs: Moving beyond 'business-as-usual'. *Sustainable Development*, *24*(6), 371–382. https://doi.org/10.1002/sd.1623

Stankov, I., Garcia, L. M. T., Mascolli, M. A., Montes, F., Meisel, J. D., Gouveia, N., et al. (2020). A systematic review of empirical and simulation studies evaluating the health impact of Transportation interventions. *Environmental Research*, *186*, 109519. https://doi.org/10.1016/j.envres.2020.109519

Statista. (2019). Smartphone penetration in India 2014–2019. *Statista*. Retrieved 12 August, 2019 from https://www.statista.com/statistics/257048/smartphone-user-penetration-in-india/

Sudmant, A., Mi, Z., Oates, L., Tian, X., & Gouldson A. (2020). *Towards sustainable mobility and improved public health: Lessons from bike sharing in Shanghai, China*. Monograph. Coalition for Urban Transitions 12, 2020. Retrieved from https://urbantransitions.global/en/publication/china-frontrunners/

Sudmant, A., Colenbrander, S., Gouldson, A., & Chilundika, N. (2017). Private opportunities, public benefits? The scope for private finance to deliver low-carbon transport Systems in Kigali, Rwanda. *Urban Climate*, *20*, 59–74.

Sudmant, A., Verlinghieri, E., Khreis, H., & Gouldson, A. (2020). Chapter 19-The social, environmental, health, and economic impacts of low carbon transport policy: A review of the evidence. In H. Khreis, M. Nieuwenhuijsen, J. Zietsman, & T. Ramani (Eds.), *Traffic-related air pollution* (pp. 471–493). Elsevier. https://doi.org/10.1016/B978-0-12-818122-5.00019-3

The Times of India. (2019). Over 1,000 electricity poles uprooted in Belagavi district, Hubballi News, *Times of India*. Retrieved 12 August 2019 from https://timesofindia.indiatimes.com/city/hubballi/over-1000-electricity-poles-uprooted-in-belagavi-dist/articleshow/64469603.cms

Too, L., & Earl, G. (2010). Public transport service quality and sustainable development: A community stakeholder perspective. *Sustainable Development*, *18*(1), 51–61. https://doi.org/10.1002/sd.412

Torres, A. B., Ortega, A. Z., Sudmant, A., & Gouldson, A. (2021) *Sustainable mobility for sustainable cities: Lessons from cycling schemes in Mexico City and Guadalajara, Mexico*. p. 45.

Tzika-Kostopoulou, D., & Nathanail, E. (2021). Exploring the big data usage in transport modelling. In E. G. Nathanail, G. Adamos, & I. Karakikes (Eds.), *Advances in intelligent systems and computing: Advances in mobility-as-a-service systems* (pp. 1117–1126). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-61075-3_107

United Nations Economic, & Social Commission for Asia and the Pacific (UNESCAP). 2014. *India Sustainable Urban Transport Program (SUTP)-BRTS Experience*. World Bank Group. Retrieved 10 July, 2019 from https://www.unescap.org/sites/default/files/4a.1_IndiaSUTP_Experience%20with%20BRT_NupurGupta.pdf

Van Crombrugge, R. (2020). The derailed promise of a participatory Minipublic: The Citizens' assembly bill in Flanders. *Journal of Deliberative Democracy*, *16*(2), 63–72. https://doi.org/10.16997/jdd.402

Venter, C., Jennings, G., Hidalgo, D., & Pineda, A. F. V. (2018). The equity impacts of bus rapid transit: A review of the evidence and implications for sustainable transport. *International Journal of Sustainable Transportation*, *12*(2), 140–152. https://doi.org/10.1080/15568318.2017.1340528

Villanueva, F. J., Aguirre, C., Rubio, A., Villa, D., Santofimia, M. J., & López, J. C. (2016). Data stream visualization framework for smart cities. *Soft Comput*, *20*, 1671–1681. https://doi.org/10.1007/s00500-015-1829-8

Wang, N., & Ma, M. (2021). Public–private partnership as a tool for sustainable development–What literatures say? *Sustainable Development*, *29*(1), 243–258. https://doi.org/10.1002/sd.2127

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—It's time to make it fair. *Nature*, *559*(7714), 324. https://doi.org/10.1038/d41586-018-05707-8

How to cite this article: Sudmant, A., Viguié, V., Lepetit, Q., Oates, L., Datey, A., Gouldson, A., & Watling, D. (2021). Fair weather forecasting? The shortcomings of big data for sustainable development, a case study from Hubballi-Dharwad, India. *Sustainable Development*, *29*(6), 1237–1248. https://doi.org/10.1002/sd.2221