Quantifying Wider Economic Impacts of Agglomeration for Transport Appraisal: Existing Evidence and Future Directions

Daniel J. Graham
Imperial College London
London, SW7 2AZ, UK

Email: d.j.graham@imperial.ac.uk

Stephen Gibbons
London School of Economics
London, WC2 2AE, UK

Email: s.gibbons@lse.ac.uk

May 2018

Executive Summary

This technical report is concerned with the Wider Economic Impacts (WEIs) of transport improvements that arise via scale economies of agglomeration. It reviews the background theory and empirical evidence on agglomeration, explains the link between transport and agglomeration, and describes a three step procedure to appraise agglomeration impacts for transport schemes within Cost Benefit Analysis (CBA). The report concludes with a set of recommendations for future empirical work on agglomeration and transport appraisal.

The main findings of the report are as follows.

- A key feature of the distribution of economic activity is a tendency towards spatial concentration, or agglomeration. We can observe this phenomenon at the level of cities, which contain vast concentrations of economic activity despite high land prices, rents and other costs. We can also observe forces of agglomeration at an industrial level, for instance in the spatial concentration of financial sectors in Wall Street or the City; or in the co-location of information technology firms around Silicon Valley.
- Agglomeration produces benefits for firms via positive external scale economies, mainly in the form of improved opportunities for sharing, matching and learning. Theory predicts that these benefits will give rise to higher productivity and lower average costs.
- Empirical work on agglomeration has tested for productivity impacts by estimating elasticities of productivity with respect to agglomeration. Estimation of a positive elasticity is viewed as consistent with the existence of agglomeration economies.
- The general consensus in the literature is that agglomeration economies exist and that they induce higher productivity for firms and workers. The unweighted mean elasticity from 47 international empirical studies is 0.046.
- Venables (2007) demonstrates an important link between transport improvements and agglomeration. Agglomeration economies can be intensified without increasing the physical concentration of firms and worker, but rather by improving *transport connectivity*. This is because the generalised costs (GCs) of travel largely determine effective economic concentration through their influence on *access to economic mass*.
- Such agglomeration effects of transport improvements are classed as Wider Economic Impacts (WEIs) because they are viewed as additional to conventional user benefits. Additionality arises from the fact that agglomeration effects are externalities induced via increasing returns to economic mass.
- To calculate WEIs that arise via agglomeration economies a three-step procedure can be used.

    1. **Calculate a connectivity metric to represent agglomeration.** Such metrics are usually based on measures of access to economic mass (ATEM), such as an effective density of the form

$$\rho_i = \frac{1}{n} \sum_{j=1}^{n} m_j f(d_{ij}),$$

where for n zones indexed by i, $i = (1, ...n)$, or j, $j = (1, ...n)$, m_j is a measure of economic mass at zone j and $f(\cdot)$, often referred to as the impedance function, is a decreasing function of the cost of travelling from origin i to destination j.

2. **Estimate elasticities of productivity (ω) with respect to a agglomeration.** Agglomeration elasticities, which we denote $\delta_s = \partial \log \omega_s / \partial \rho$, are estimated separately for each industrial sector s, $s = (1, ...S)$, using econometric models of the form

$$\omega_{si} = f(\rho_i, Z_{si}),$$

where Z_{si} represents other relevant effects on productivity.

3. **Quantify the agglomeration impacts arising from transport schemes using ρ and δ_s.** This involves calculating the expected change in effective density, and then re-scaling this to give an expected change in productivity using the δ_s estimates.

- The report reviews different approaches that can be use to implement these three steps. The key points are as follows.

1. **Connectivity metrics** - a variety of different mass and impedance measures can be used to calculate connectivity metrics. For example, population or employment (or the sum of both) could be used as measures of mass; while travel times, or generalised cost by mode, or some other measure of the resistance of travel can be used as impedance functions. Another important issue discussed in the report concerns how the spatial *decay* of agglomeration effects should be represented.
2. **Estimation of agglomeration elasticities** - productivity can be represented via TFP within a production function framework or via labour productivity with a model for wages assuming that workers are paid the value of their marginal product. In adopting either approach there are econometric challenges that must be addressed in order to obtain valid causal evidence on the effect of agglomeration on productivity.
3. **Calculating productivity effects of transport schemes** - the report discusses a number of different approaches that could be used to calculate the agglomeration effects of transport schemes and provides an evaluation of each. It also considers whether appraisal calculations could distinguish between urbanisation and localisation agglomeration effects, and whether the effect of agglomeration on productivity should be assumed constant for all locations or allowed to vary.

- The report concludes with the following recommendations for future work.

R1 Estimation of agglomeration elasticities should be conducted using different measures of mass and impedance for MED variables, including use of population and employment to represent mass and average GC as well as distance to represent impedance.

R2 Alternative approaches to estimate the distance decay of agglomeration should be implemented and compared and the implications for appraisal calculations evaluated.

R3 Agglomeration elasticities should be estimated via both wage and TFP models, using consistent measures of ATEM for the same spatial units over the same time period. Results should be compared and a judgement made as to which evidence is most robust and suitable for use in appraisal.

R4 Different econometric models using different covariate specifications should be tested to observe the robustness of elasticity estimates to model assumptions (e.g. conditional versus unconditional estimates).

R5 Due to limitations of existing data, and econometric challenges arising from severe multicollinearity, we do not recommend that attempts be made to estimate mode specific agglomeration elasticities.

R6 Models that distinguish localisation and urbanisation effects should be estimated with a view to deciding whether the resulting evidence is suitable for use in appraisal.

R7 Econometric models should be designed to explore the existence of heterogeneous agglomeration effects by allowing for nonlinearities and by estimating separate agglomeration elasticities for sub-samples of the data based on area type. Consideration should be given as to whether the resulting evidence is suitable for use in appraisal.

1 Introduction

Cost Benefit Analysis (CBA) uses concepts from economic theory to measure the change in net 'social-welfare' arising from transport improvements. An increase in social welfare occurs when the benefits that accrue to society are greater than the costs. In CBA, benefits and costs are calculated in monetary values, largely by approximating change in consumers' surplus. Summary measures of value for money are then produced such as the net present value of the scheme and the benefit cost ratio (BCR). CBA forms a key component of ex-ante project appraisal in the UK (for a recent review of CBA see Mackie et al. 2012).

CBA has a well established theoretical and empirical basis and it provides a familiar and well understood approach that is routinely used by Civil Servants, transport professionals, and academics. It has been recognised for some time that the conventional consumer surplus based calculation of conventional CBA capture only a sub-set of the potential benefits of transport schemes. Recent work on Wider Economic Impacts (WEIs) has extended the scope of appraisal to incorporate impacts arising from externalities and from forms of imperfect competition, again based on clearly set out theoretical and empirical evidence.

In this technical report we discuss calculation of the WEIs of transport improvements that arise via scale economies of agglomeration. The report is structured as follows. Section 2 briefly reviews the background theory and empirical evidence on agglomeration. Section 3 explains the link between transport and agglomeration and outlines a three step procedure to appraise agglomeration impacts within CBA. These three steps are then discussed in detail in sections 4, 5 and 6. The final section of the report presents recommendations for the next phase of work.

2 Urban agglomeration economies

A key feature of the distribution of economic activity is a tendency towards spatial concentration, or agglomeration. We can observe this phenomenon at the level of cities, which contain vast concentrations of economic activity despite high land prices, rents and other costs. We can also observe forces of agglomeration at an industrial level, for instance in the spatial concentration of financial sectors in Wall Street or the City of London; or in the co-location of information technology firms found around Silicon Valley. Economic theory states that both forms of agglomeration are driven by spatial externalities, or what are termed agglomeration economies. Economies of industry concentration, or localisation economies, are external to the firm but internal to the industry. Economies of urban concentration, or urbanisation economies, are external to the firm and the industry but internal to the city.

Duranton and Puga (2004) discuss the micro-foundations of agglomeration and show that these are mainly driven by three simple mechanisms: sharing, matching and learning. Thus for firms, the main benefits of agglomeration arise through improved opportunities for labour market pooling, knowledge interactions, specialisation, and the sharing of inputs and outputs. The key point is that benefits accrue to firms in cities via positive external scale economies and theory predicts that these benefits will be manifest in higher productivity and lower average costs.

Accordingly, empirical work on agglomeration has sought to estimate the relationship between city size and productivity. Evidence of a positive relationship is viewed as consistent with the existence of agglomeration economies. Agglomeration has typically been measured by city size (via population or employment) or via a variable measuring the degree of access to economic mass. Productivity has been represented by wages or by Total Factor Productivity (TFP).

Table 1 reports results from 47 international empirical studies that have estimated the effects of agglomeration on productivity. The table shows the number of elasticity estimates collected from each study, the mean elasticity value, the median elasticity value, and the range of estimated elasticity values. Estimates vary between -0.800 and 0.658, and have unweighted mean equal to 0.046 and median equal to 0.043. Figure 1 provides a histogram of the values shown in the table.

The general consensus in the literature is that agglomeration economies exist and that they induce higher productivity for firms and workers, but there are differences in estimates of the magnitude of this effect.

Melo et al. (2009) conduct a meta-analysis of the empirical literature on urban agglomeration economies to investigate the influence of study-design characteristics on results. They find large differences in the size of elasticity estimates across countries reflecting differences in the nature of economies and urban systems. They also find substantial differences in the magnitude of agglomeration economies across industry sectors, with service industries tending to derive considerably larger benefits from urban agglomeration than manufacturing.

In addition to these broad contextual factors, the methodological approaches used to estimate elasticities can also have a large influence on results. This is evident both between and within studies. In particular, the magnitude of agglomeration estimates is strongly influenced by the manner in which studies have, or have not, attempted to correct for potential sources of 'endogeneity'.

Table 1: International estimates of urban agglomeration elasticities

study	*country*	*period*	*data*	*aggregation*	*obs.*	*mean*	*median*	*Range*
Aaberg (1973)	Sweden	1965-68	CS	regions	4	0.017	0.018	[0.014, 0.019]
Ahlfeldt et al. (2015)	Germany	1936-1986-2006	PD	regions	3	0.062	0.066	[0.045, 0.074]
Au and Henderson (2006)	China	1997	CS	regions	2	0.013	0.013	[-0.007, 0.033]
Baldwin et al. (2007)	Canada	1999	CS	plant	8	0.061	0.071	[-0.008, 0.104]
Baldwin et al. (2008)	Canada	1989-1999	PD	plant	6	-0.088	-0.130	[-0.310, 0.300]
Brulhart and Mathys (2008)	Europe	1980-2003	PD	regions	14	-0.080	0.055	[-0.800, 0.280]
Ciccone (2002)	Europe	1992	CS	regions	7	0.047	0.045	[0.044, 0.051]
Ciccone and Hall (1996)	US	1988	CS	regions	8	0.053	0.049	[0.035, 0.084]
Cingano and Schivardi (2004)	Italy	1992	CS	regions	13	0.054	0.064	[0.019, 0.073]
Combes et al. (2010)	France	1988	PD	worker	43	0.035	0.037	[0.012, 0.054]
Combes et al. (2008)	France	1988	PD	zone	11	0.052	0.035	[0.024, 0.143]
Combes et al. (2012)	France	1994-2002	PD	plant	17	0.090	0.070	[0.040, 0.190]
Davis and Weinstein (2003)	Japan	1985	CS	regions	11	0.027	0.028	[0.010, 0.057]
DiAddario and Patacchini (2008)	Italy	1995-2002	PD	worker	1	0.010	0.010	[0.010, 0.010]
Fingleton (2003)	UK/GB	1999-2000	CS	regions	3	0.017	0.016	[0.016, 0.018]
Fingleton (2006)	UK/GB	2000	CS	regions	7	0.025	0.018	[0.014, 0.049]
Graham (2000)	UK/GB	1984	CS	regions	22	-0.006	-0.001	[-0.168, 0.141]
Graham (2005)	UK/GB	1998-2002	PD	firm	36	0.193	0.171	[-0.037, 0.503]
Graham (2007b)	UK/GB	1995-2004	PD	firm	28	0.110	0.098	[-0.191, 0.382]
Graham (2007a)	UK/GB	1995-2004	PD	firm	18	0.194	0.195	[0.041, 0.399]
Graham (2009)	UK/GB	1995-2004	PD	firm	108	0.097	0.083	[-0.277, 0.491]
Graham and Kim (2008)	UK/GB	1995-2004	PD	firm	18	0.079	0.049	[-0.13, 0.306]
Graham et al. (2009)	UK/GB	2000-2006	PD	plant	5	0.041	0.034	[0.021, 0.083]
Graham and Van Dender (2011)	UK/GB	1995-2004	PD	firm	6	0.072	0.061	[0.009, 0.134]
Henderson (1986)	Brazil	1970-72	CS	regions	52	0.010	0.018	[-0.366, 0.18]
Henderson (2003)	US	1982	PD	firm	4	0.024	0.017	[-0.127, 0.189]
Hensher et al. (2012)	Australia	2006	CS	zone	39	0.071	0.051	[-0.049,406]
Holl (2012)	Spain	1991-2005	PD	firm	23	0.089	0.047	[-0.079, 0.827]
Kanemoto et al. (1996)	Japan	1985	CS	regions	9	0.089	0.070	[0.010, 0.250]
Lall et al. (2004)	India	1991	CS	plant	18	0.017	0.007	[-0.204, 0.658]
Mare (2016)	NZ	2001-2012	PD	plant	31	0.075	0.075	[0.0405, 0.116]
Mare and Graham (2013)	NZ	1999-2007	PD	plant	114	0.043	0.048	[-0.13, 0.222]
Marrocu et al. (2013)	Europe	1996-2007	CS	regions	5	0.036	0.041	[0.027, 0.040]
Martin et al. (2011)	France	1996-2004	PD	plant	8	0.011	0.010	[-0.06, 0.066]
Melo and Graham (2009)	UK/GB	2002-2006	PD	worker	64	0.029	0.020	[-0.13, 0.114]
Mion and Naticchioni (2005)	Italy	1995	PD	worker	30	0.034	0.022	[0.002, 0.109]
Moomaw (1981)	US	1967	CS	regions	18	0.060	0.032	[0.006, 0.319]
Moomaw (1983)	US	1977	CS	regions	26	0.038	0.034	[-0.052, 0.182]
Moomaw (1985)	US	1972	PD	regions	36	0.040	0.036	[-0.104, 0.27]
Morikawa (2011)	Japan	2002-2005	PD	firm	4	0.110	0.110	[0.070, 0.150]
Nakamura (1985)	Japan	1979	CS	cities	38	0.026	0.022	[-0.037, 0.081]
Rice et al. (2006)	UK/GB	1998-2000	CS	regions	14	0.026	0.024	[-0.005, 0.07]
Rosenthal and Strange (2008)	US	2000	CS	worker	9	0.042	0.046	[0.025, 0.058]
Sveikauskas et al. (1988)	US	1977	CS	regions	6	0.013	0.014	[0.007, 0.017]
Sveikauskas (1975)	US	1967	CS	regions	42	0.057	0.054	[0.012, 0.124]
Tabuchi (1986)	Japan	1980	CS	regions	57	0.060	0.056	[-0.079, 0.300]
Wheeler (2001)	US	1980	CS	worker	3	0.017	0.020	[0.000, 0.030]
Average					1043	0.046	0.043	[-0.800, 0.658]

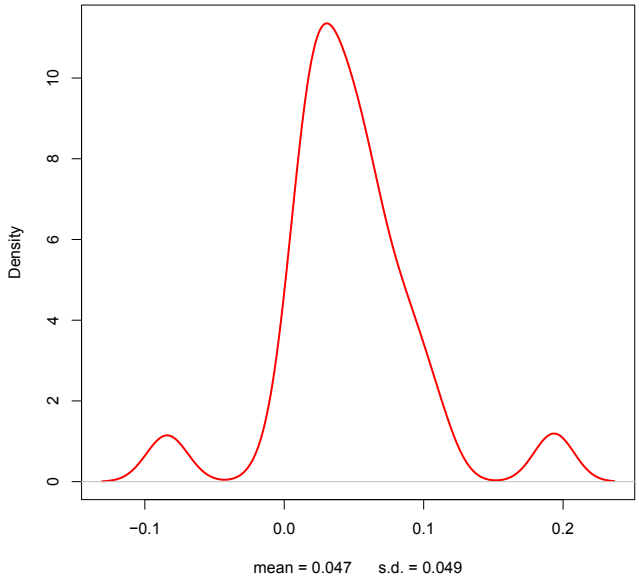

Figure 1: Histogram of urban agglomeration elasticities.

In summary, there is a great deal of empirical evidence indicating that productivity is higher in cities, and this is consistent with the theory of urban agglomeration. The data and methods used to estimate the relationship between agglomeration and productivity matter for the results obtained, and the recent literature has made substantial progress in understanding the conditions required for valid inference.

3 Appraising agglomeration impacts within CBA: current theory and practice

3.1 Direct and wider impacts of transport improvements

Current CBA practice in the UK classifies the benefits of transport improvements under two broad headings (for details see Mackie et al. 2012, Venables et al. 2014, DfT 2014).

1. **Direct user-benefits (DUBs)** - DUBs capture impacts that are generated for both new and existing users of the transport system. They arise via changes in the generalised cost of travel (e.g. in time and operating costs) or in the quality of transport services. DUBs typically constitute the largest component of benefits within conventional CBA calculations. The economic theory underpinning CBA shows that under conditions of perfect competition, constant returns to scale, and in the absence of any market failures; all economic impacts of transport schemes would be captured via DUBs. These idealised economic conditions, however, are never met in practice and as a result there has long been a recognition that DUBs capture only a sub-set of potential economic impacts.
2. **Wider economic Impacts (WEIs)** - WEIs refer to impacts on the economy that arise via market failures and are therefore described as wider, or additional to conventional user benefits. The UK CBA guidance defines three types of WEIs
    i. **Agglomeration economies** - transport improvements can increase the scale of potential economic interactions available in the economy, with implications for the relative level of agglomeration experienced by firms.
    ii. **Imperfect competition** - transport improvements can cause a decrease in the costs of interacting in the spatial economy, thus potentially allowing firms to expand output. Output expansion yields a welfare gain in monopolistic markets when willingness to pay for the increased output exceeds the cost of producing it.
    iii. **Tax revenues arising from labour market impacts** - the decisions that firms and workers make about where to locate is influenced by the accessibility offered by transport systems. If accessibility improves, and causes firms / workers to move to more productive locations or have greater participation in labour markets, this will result in a tangible financial gain (i.e. higher wages or productivity). Most of this is captured in the consumer surplus based calculations of user benefits, but not the resulting change in tax revenue to the government (i.e. income tax, national insurance, and corporation tax).

It is important to stress that WEIs are viewed as additional to DUBs because they derive from sources of market failure and imperfect competition. The actual reduction in transaction costs brought about by a transport improvement provides a direct benefit to consumers and producers that is captured under the conventional CBA DUB calculations. It is only the 'wider' implications of this fall in transaction costs, for scale economies of agglomeration or for spatial competition, that forms the additional WEI component.

3.2 Transport improvements and agglomeration

There is an inherent relationship between transport and the externalities of agglomeration. For a fixed distribution of firms and people across space, transport improvements reduce the costs of interaction, between firms, between workers, between workers and firms, and between firms and consumers. These are referred to as the 'static' agglomeration effects, since firms and workers are not changing workplace or residential location or becoming more spatially concentrated; it is only the costs of agents interacting across locations that changes. Transport improvements also make some places more attactive than others as places to live and work, which can lead to firm and worker relocations, land use change and - potentially - additional agglomeration effects due to increased concentration of agents. These are often referred to as 'dynamic' agglomeration effects or 'clustering'. The static agglomeration effects result from changes in the 'effective' density of firms and worker, even if the actual spatial density is unchanged.

From this line of reasoning it is clear that there may be consequences of transport investment that relate specifically to agglomeration. Essentially, the argument is that if there are increasing returns to effective spatial concentration, and if transport in part determines the level of effective concentration or density experienced by firms, then investment in transport may induce some shift in the productivity of firms via externalities of agglomeration.

Venables (2007) develops a model of the relationship between transport and agglomeration that illustrates three important points. First, he shows in what way transport improvements can generate productivity benefits via agglomerations economies. Second, he shows that such benefits are genuinely additional to the DUBs of conventional CBA . Third, he suggests a way of making simple calculations to quantify agglomeration impacts within the standard CBA framework.

The Venables (2007) model is basically a combination of the Henderson (1974) city size model and the standard Alonso-Muth-Mills monocentric city model. In his exposition he considers an intra-urban transport improvement which reduces the costs of commuting to the centre of the city. The edge of the city commuting area is defined by the distance or city population at which commuting costs offset the wage premium offered by central city jobs. When commuting costs are reduced, people outside the edge of the city formerly working outside it, now find it worthwhile to commute to higher paid central city jobs, increasing employment in the central city. Central city wages increase as a result, if there are productivity effects associated with higher city employment. His analysis delineates a) the cost savings arising from the transport improvements, for both existing and new commuters, which are the user benefits in the CBA calculations; and b) the agglomeration-related productivity effects due to increased city employment, which generate wage gains for all workers, which are the 'wider impacts' not captured in traditional CBA calculations.

A diagramatic representation of Venables' model is reproduced as figure 2.

The horizontal axes in the figure represent city size while the vertical axes measure costs and benefits. The wage gap represents the difference between urban and non-urban wages. In the absence of agglomeration economies, shown in the left hand diagram, an urban equilibrium is initially established at city size X, where the wage gap between city workers and non-city workers is entirely dissipated in the travel costs of the city worker that is most distant from the CBD.

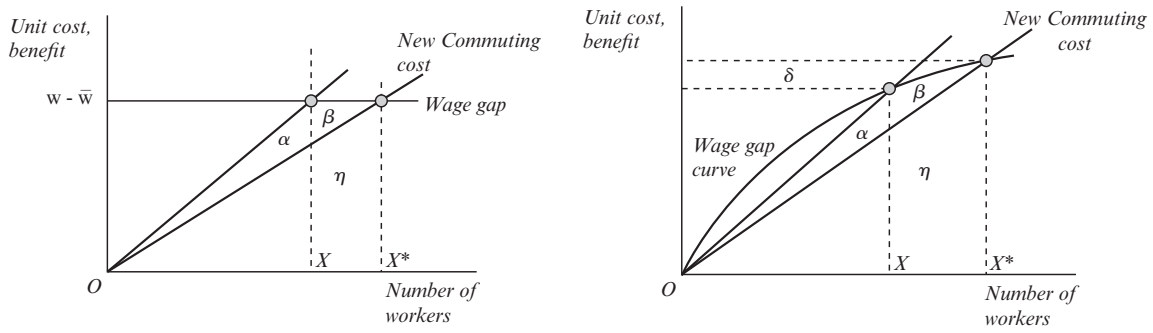

Figure 2: Net gains from a transport improvement without and with agglomeration effects.

When a transport improvement is made, commuting costs are shifted downwards and consequently the city expands to point X^*. The change in output $(\beta+\eta)$ minus the total change in the resources used in commuting is $\eta-\alpha$, yields a net benefit from the transport improvement of $\alpha+\beta$. This is the DUB captured in standard CBA consumer surplus calculations.

The diagram on the right-hand sides assumes that agglomeration (urbanisation) economies exist, and thus we see that productivity (or wages) are increasing with city size as reflected in the concave wage gap curve. Equilibrium is again found at the intersection of the commuting cost and wage gap curves, but the fact that productivity is non-constant with respect to city size means that the real income gain from a transport improvement is $\alpha+\beta+\delta$; where δ measures the increase in productivity experienced by city workers and is the total increase in productivity per worker due to the increase in city size. The slope of the wage curve is the derivative of wages (or labour productivity) with respect to city size. If the wages and city size were measured in natural logs, the slope would be the elasticity of wages (or productivity) with respect to city size, which is the elasticity which the empirical work discussed above aims to estimate. Note that δ captures an *externality* which by definition is excluded under the assumptions used to make the consumers surplus based calculations of CBA.

In this way Venables demonstrates that there can be external benefits from transport investment related to agglomeration and that these are additional to conventional user benefits.

Although the Venables model is concerned with the role of transport in increasing the effective commuting area of a city, and the number of workers in the city centre, the same principles have been extended to think more generally about the static agglomeration effects induced by transport improvements. For example, we can think of the CBD in the diagram as any workplace location, and the x axis as agents (workers, firms) arranged uniformly over distance from that location. A reduction in transport costs per unit of distance increases the number of agents that can reach that location before the costs offset the benefits, or, equivalently, the 'effective density' of agents measured by the number of agents accessible per unit of time or cost. If there are agglomeration effects due to the number of agents that can access the location, then the transport cost reductions induce productivity gains. It is this idea of the productivity benefits from increasing the effective density of places that has worked its way into the DfT's transport appraisal guidance.

These agglomeration impacts can be quantified quite simply if we know : a) the change in effective density that will result from making some transport intervention; and, b) the amount

by which productivity will rise in response to an increase in effective density.

The direct user benefits and wider impacts seem conceptually clear in this diagrammatic analysis - the user benefits are due to transport cost reductions for transport users, the wider impacts are externalities from interaction between agents that increases their productivity. However, questions remain about the extent to which the user and wider impacts are completely separable in practice, because of the way that the value of transport costs reductions to users and the productivity benefits to firms or wage benefits to workers are inferred empirically, and ambiguities in their interpretation. At the core of the problem are concerns about: a) on the transport demand side, the extent to which the agglomeration benefits are already being priced in to the 'value of travel time savings' which form the basis for estimating the user benefits; b) on the production side, the extent to which transport cost reductions affect productivity and show up in the estimated relationship between density and productivity. The first of these questions about 'additionality' is discussed at length in a DfT report (DfT 2007). The second has received less attention.

On the transport demand side, the traditional approach to evaluating the user benefits (consumer surplus) is based on a 'value of travel time savings'. This figure represents the rate at which individuals or employers are willing to trade off time travelling against cost when making transport mode or route decisions. It is a number that is derived empirically, from stated preference, experimental evidence or revealed preferences. Put simply, the basic user benefits from a transport scheme are simply the value of the time savings aggregated over existing users with an adjustment for the new users due to induced demand (the time savings per user $\times$ half the number of predicted new users). These user benefits correspond to the value of the additional input of worker time available for production (or in leisure) due to the time savings arising from the relaxation of a constraint on travel time due to the transport improvement.

Take a setting where values of travel time savings are elicited from a situation where an individual is making a judgement between alternative modes or routes with different journey times and costs. For the values of travel time savings to represent only the user benefits, individuals must be assumed to be making this decision with their current wage per hour and hence productivity per hour in mind and ignoring the impact that their decision has on other individuals in the economy. While a person making this internal calculation may take into account that the travel time saving means they can work an additional hour and earn an hour's additional wage, there is no presumption that they will take into account any additional productivity per hour or higher hourly wage that that the time saving might induce. In particular, they explicitly will not take into account the increases in the wage induced by greater general connectivity and the externalities that underlie the notion of 'agglomeration economies' discussed above. These are the externalities arising from improved connectivity and greater interaction between multiple agents over the network, not the transport choices of a single agents. Moreover, time freed up from reduced travel time may not be used for productive activities at all, but may be transferred to leisure time, with no productivity impacts. A similar distinction is made by Venables et al. (2014) who emphasise the distinction between the welfare improving aspects of transport improvements and those that involve increases in GDP, value-added or productivity.

Note, therefore, that in the above scenario, the value of time savings does not relate at all to productivity improvements: it captures the value of time savings holding wages and productivity constant. However, there are exceptions. An obvious exception is the case of

travel that takes place during hours in which a worker is counted as employed. This would include, for example, some business travel and driver hours in freight transportation. In such cases, employers' willingness to pay for reduced travel times might reflect the increased productivity of their employees, when output is measured as output per worker per year or output per contracted hour of work, or the hourly wage (assuming wages correspond to worker productivity).

On the production side, the additionality question hinges on whether estimates of the relationship between effective density and productivity partly capture some of the user benefits discussed above. Productivity, in this context, can mean the quantity of output produced for a fixed index of measurable inputs (Total Factor Productivity), the quantity of output produced per unit of labour input (labour productivity) or a proxy such as a wage rate. The details on what inputs are accunted for in this calculation are potentially important. In a situation where we are measuring productivity per worker, or productivity per firm, it is conceivable that in places where effective density is high and transport costs low, part of the estimated 'productivity' advantage may be due to workers supplying more hours, which has nothing to do with the externalities associated with agent interactions, but due to the workers allocation of time in dense/low transport cost places relative to sparse/high transport cost places - which is in turn part of the user benefits. Another reason for some overlap between the user and wider impacts on the production side could be that transport cost reductions or closer proximity between firms reduces the costs of intermediate goods inputs. Reductions in the costs of intermediate goods inputs will show up in the data as an increase in measured value-added and hence potentially in productivity. Again, in this scenario, estimates of higher productivity in high density/low transport cost areas would reflect lower cost inputs, not an externality. Ideally, to ensure that density-productivity elasticity estimates capture only the externalities, we would want to control completely for labour hours, input prices, output prices and a wide range of other inputs, although this is rarely possible given the data available. We say more about appropriate methods and data to estimate the agglomeration elasticities later in this report.

This distinction between productivity effects and time-allocation related saving is important to the additionality of the wider benefits calculations, given that the parameters that are used in the wider benefits calculations are estimated from differences in wages or productivity between workers in different places. Unfortunately, as the above discussion suggests, it is impossible in general to provide a clear-cut partitioning of the empirical estimates of values of travel time savings and the productivity improvements from greater effective density into overlapping and non-overlapping components. The exact details will depend very much on how values of time and the productivity-density elasticities have been estimated.

3.3 Steps to calculate wider economic benefits of agglomeration

To calculate WEIs that arise via agglomeration economies the following three-step procedure is required.

1. **Calculate access to economic mass (ATEM) via effective densities**

   There are n zones indexed by i, $i = (1, ...n)$, or j, $j = (1, ...n)$. The Mean Effective

Density (MED) for zone i, which we denote ρ_i is calculated as

$$\rho_i = \frac{1}{n} \sum_{j=1}^{n} m_j f(d_{ij}).$$

where m_j is some measure of economic mass at zone j and $f(\cdot)$, often referred to as the impedance function, is a decreasing function of the cost of travelling from origin i to destination j. Cost could be a function of distance, travel time, generalised cost or some other indicator of the resistance of travel.

The MED measure of ATEM is deigned to

- capture the effects of both scale and spatial proximity
- provide a flexible spatial framework largely free of arbitrary boundaries
- incorporate an implicit transport accessibility dimension via the impedance function

The MED calculations should be made for small spatial zones of the city or region of interest.

2. Estimate agglomeration elasticities

To estimate an agglomeration elasticity, which we denote δ, for some economic sector the following general procedure can be used.

i. Obtain data on the distribution of productivity across space. This is normally done using spatially referenced firm level production data, on output and inputs, or spatially referenced wage data.

ii. Using the firm or worker data, map the location of each 'unit' (i.e. firm or worker) using GIS and superimpose a zone GIS layer for calculation of MED measures.

iii. Calculate MED measures of agglomeration for each zone and assign the zone value to units within each zone.

iv. Specify a production function, or wage equation, with the MED agglomeration variable included as a shifter of productivity, i.e.

$$y_i = g(\rho_i) f(x_i),$$

where y_i represents output and x_i a vector of factor inputs for zone i; or

$$w_i = g(\rho_i) f(z_i)$$

where w_i is the wage rate and z_i is a vector of covariates relevant for wage determination

v. Obtain estimates of $\delta^s = \partial \log y^s / \partial \log \rho_i$, or $\delta^s = \partial \log w_i / \partial \log \rho_i$ separately for each industrial sector s, $s = (1, ..., S)$, of interest.

It is important to note that using observed data we can construct only a partial representation of sources of productivity. Consequently, there are a number of estimation issues that need to be carefully addressed to ensure that the agglomeration elasticity estimates are, as far as possible, causal rather than simply associational. These issues are discussed later in the paper in the section on econometric methods.

3. Quantify the agglomeration benefits arising from transport schemes

Quantifying the agglomeration benefits of a transport scheme requires a calculation of how the proposed transport scheme is expected to change productivity. This involves calculating the expected change in effective density, and then re-scaling this to give an expected change in productivity using the δ^s estimates as described above.

The first step is to calculate how the proposed transport scheme under consideration will change the effective density. The MED for a particular unit in the initial period is $\rho_i^0 = \sum m_j f(d_{ij}^0)$ and in the post-improvement period is $\rho_i^1 = \sum m_j f(d_{ij}^1)$, holding everything else constant apart from the changes in the travel costs: $f(d_{ij}^1) - f(d_{ij}^0)$. Note that in this calculation for typical travel schemes, distance is an inappropriate metric for the change in impedance since transport improvements rarely primarily entail reductions in distance. So even if the agglomeration elasticity has been estimated using an MED measure based on minimum straight line distances between zones, the actual effective density changes of the transport improvement are derived by considering the proposed change in generalised costs or times. We discuss this point in more detail below.

These changes in travel times or generalised transport costs along all OD pairs $i-j$ in the network must be derived from a transport model, or otherwise inferred from information on the expected change in minimum journey times or costs across the whole network. Note that an improved direct link between any two zones will potentially affect other zones, since the new link may offer a new quicker journey time between other pairs.

With this change in effective density in hand, the next step is to predict the change in productivity from the elasticity estimate, δ, described above. First consider the case where one elasticity, δ, has been estimated across all industries. The percentage productivity change for an individual spatial unit i is simply the percentage change (or difference in logs) in effective density in unit i multiplied by the elasticity of productivity with respect to effective density (δ). Since $\mathbb{E}[\Delta \log y_i] = \delta \mathbb{E}[\Delta \log \rho_i]$, the change in each spatial unit is predicted as $\Delta \log y_i = \delta \Delta \log \rho_i$. The unit specific changes in productivity then have to be transformed from percentage changes to monetary units and aggregated across spatial units for the area under consideration (a city, region, the whole nation) to give the total benefits.

If agglomeration-productivity elasticities have been estimated for separate industrial sectors, industry specific aggregate benefits can be calculated in monetary value via

$$\sum_{s=1}^{S}\sum_{i=1}^{n} \Delta y_i^s(\Delta\rho_i) = \sum_{s=1}^{S}\sum_{i=1}^{n} \left[y_i^s(\rho_i^1) - y_i^s(\rho_i^0)\right] = \sum_{s=1}^{S}\sum_{i=1}^{n} \left[\left(\frac{\rho_i^1}{\rho_i^0}\right)^{\delta^s} - 1\right] y_i^s(\rho_i^0)$$

where $y_i^s(\rho_i^0)$ is a measure of economic output (i.e. GDP) for sector s in zone i the base scenario, ρ_i^0 is the MED measure of agglomeration for zone i in the base scenario and ρ_i^1 is the predicted value of agglomeration after the transport scheme is in place. This is the calculation suggested in WebTag and currently used in UK CBA practice.

In the following three sections of the report we go through each of the three calculation steps in turn, providing detail on the different ways in which they can be implemented.

4 Measuring Access to Economic Mass (ATEM): Effective Densities

The first step in the agglomeration WEI calculation outlined in section 3 constructs a mean 'effective density' (MED) measure of access to economic mass (ATEM) to represent agglomeration. In this section of the report we show how MEDs represent ATEM and we comment on the different functional forms and sources of data that can be used to construct them.

4.1 Decomposition of Effective Densities

It is useful to decompose MEDs to reveal more about the forces that they represent. In doing so we will consider MEDs based on Euclidean distance of the form

$$\rho_i^D = \frac{1}{n}\sum_{j=1}^{n}\frac{m_j}{d_{ij}^{\alpha}},$$

which was that used to estimate the agglomeration elasticities currently recommended for use in WebTag. Applying a law of large numbers to ρ_i^D we have

$$\rho_i^D \xrightarrow{p} \mathbb{E}\left(M_j D_{ij}^{-\alpha}\right) = \mathbb{E}(M_j)\mathbb{E}\left(D_{ij}^{-\alpha}\right) + \mathrm{Cov}\left(M_j, D_{ij}^{-\alpha}\right). \tag{1}$$

The first term on the right-hand side of (1), which we refer to as scaled centrality (SC), is the product of the mean mass and mean impedance function values. The former is constant across zones but the latter varies according two key factors.

1. **Centrality** - the value of $\mathbb{E}(D_{ij}^{-\alpha})$ depends on the geographic location of the zone, such that, other things being equal, peripheral zones will have lower values and core zones higher values. This effect is illustrated in figure 3 for zones of constant size and mass formed as a uniform grid, and with α set equal to 1.0. Since the covariance term must be zero with uniform mass, the figure illustrates variance in MED values that is entirely due to centrality.
2. **Zone size distribution** - the value of $\mathbb{E}(D_{ij}^{-\alpha})$ also depends on the distribution of zone sizes. This is illustrated in figure 4 which shows MED values for zones that are again of equal mass but this time of different sizes. Note that this causes higher MED values to be observed where the Euclidean distances between zone centroids are smaller.

An important observation about the SC term $\mathbb{E}(M_j)\mathbb{E}\left(D_{ij}^{-\alpha}\right)$, which we return to below, is that values derived for any two different mass measures will be perfectly correlated. This is because $\mathbb{E}(M_j)$ is simply a constant: the average mass in each spatial unit in the entire economy.

The second term on the right-hand side of (1) is the covariance of the mass and the impedance function variables. We refer to this term as the mass-impedance covariance (MIC). Clearly, with a uniform or random distribution of mass across zones the MIC will be zero. However, in real world data we typically find that the mass and distance vectors are correlated due to the

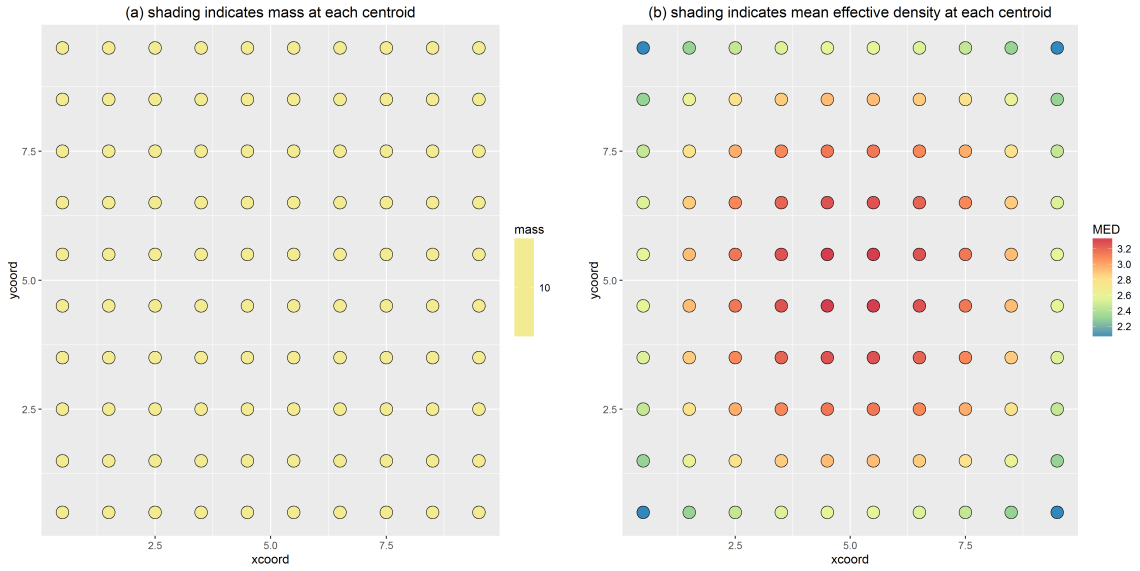

Figure 3: MED values for zones of a constant size formed as a uniform grid.

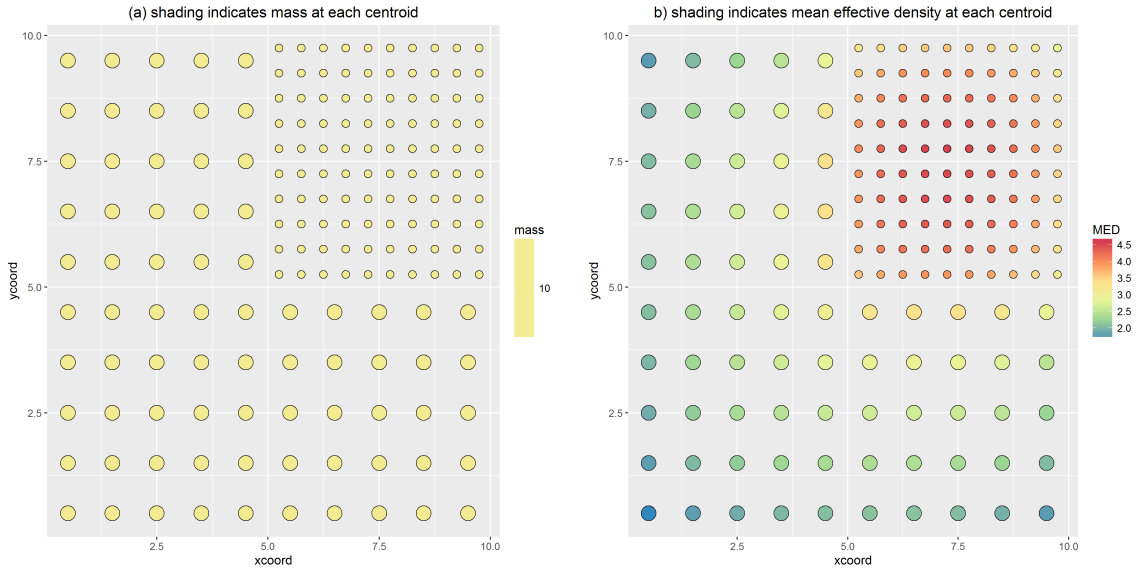

Figure 4: MED values for zones of variable size.

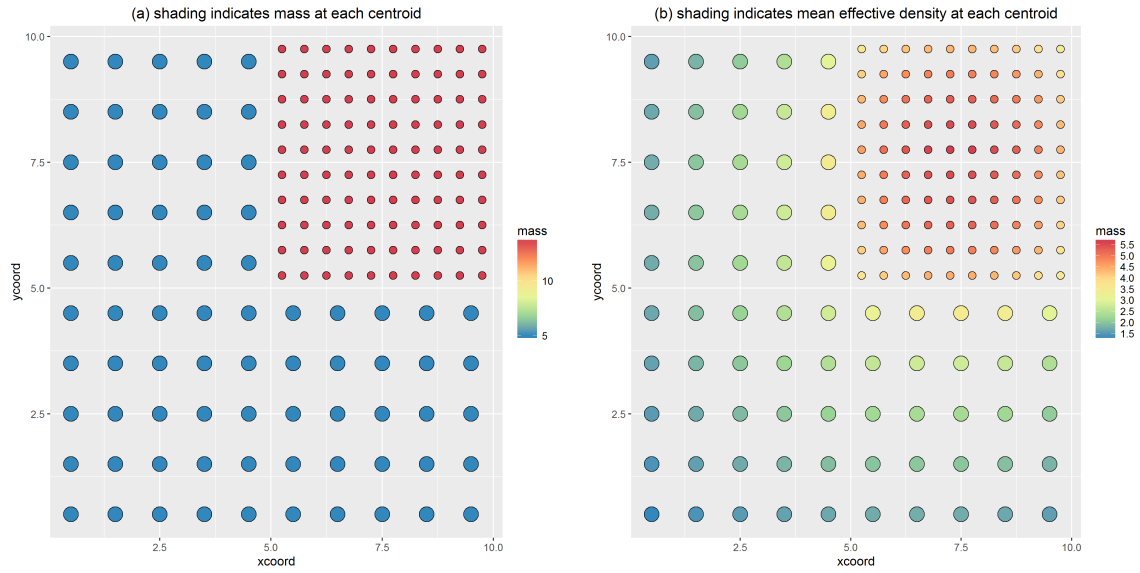

Figure 5: MED values for zones of variable size and mass.

positive influence of accessibility on agglomeration, and also because mass is often a relevant factor determining administrative zone sizes.

The consequences of an uneven mass distribution are shown in figure 5. The mean mass is the same as in previous plots, but large zones have a mass value of 5 and small zones 13.5. We can observe that the range of MED values is larger than in figure 4 with small zones having higher values due to their larger mass. Note that the MED values in figure 5 differ from those in figure 4 due solely to non-zero covariance terms, the values of $\mathbb{E}(M_j)\mathbb{E}\left(D_{ij}^{-\alpha}\right)$ are of course identical.

In summary MED values are determined by three typically interrelated factors: geographic centrality of the zones (e.g. core / periphery), size distribution of the zones, and the spatial distribution of economic mass.

Figure 6 illustrates the components of MEDs for 8480 small zones of Britain calculated using total employment as the mass variable and inverse distance as the impedance function. The employment data for England and Wales are for Middle Layer Super Output Areas (MSOAs), and for Scotland, for Intermediate Zones (IZs). The top left-hand panel shows smooth histograms of MED values and its two components (SC and MIC) (i.e. $E(M) \times E(D^{-1})$ and $\mathrm{Cov}(M, D^{-1})$). The maps plot each of the two components as well the MED itself.

The top left-hand panel of figure 6 shows densities for the MED values and their components. All three densities have a long right-hand skews reflecting high values for central city locations, and in particular, for central London. Note that the MIC component has a large range of values, from -11 to 128. High values correspond to places where zones that are closely spaced also tend to have high mass, such as in central London. Near-zero values correspond to places where there is no systematic relationship between the distance between zones and the employment in those zones. The top left panel show that most values of the MIC terms are concentrated close to zero and that the distribution of MED closely follows the distribution of SC. The MIC term contributes mainly to the right hand tail of the distribution of MED and the high MED spikes visible in central city locations in Figure 6.

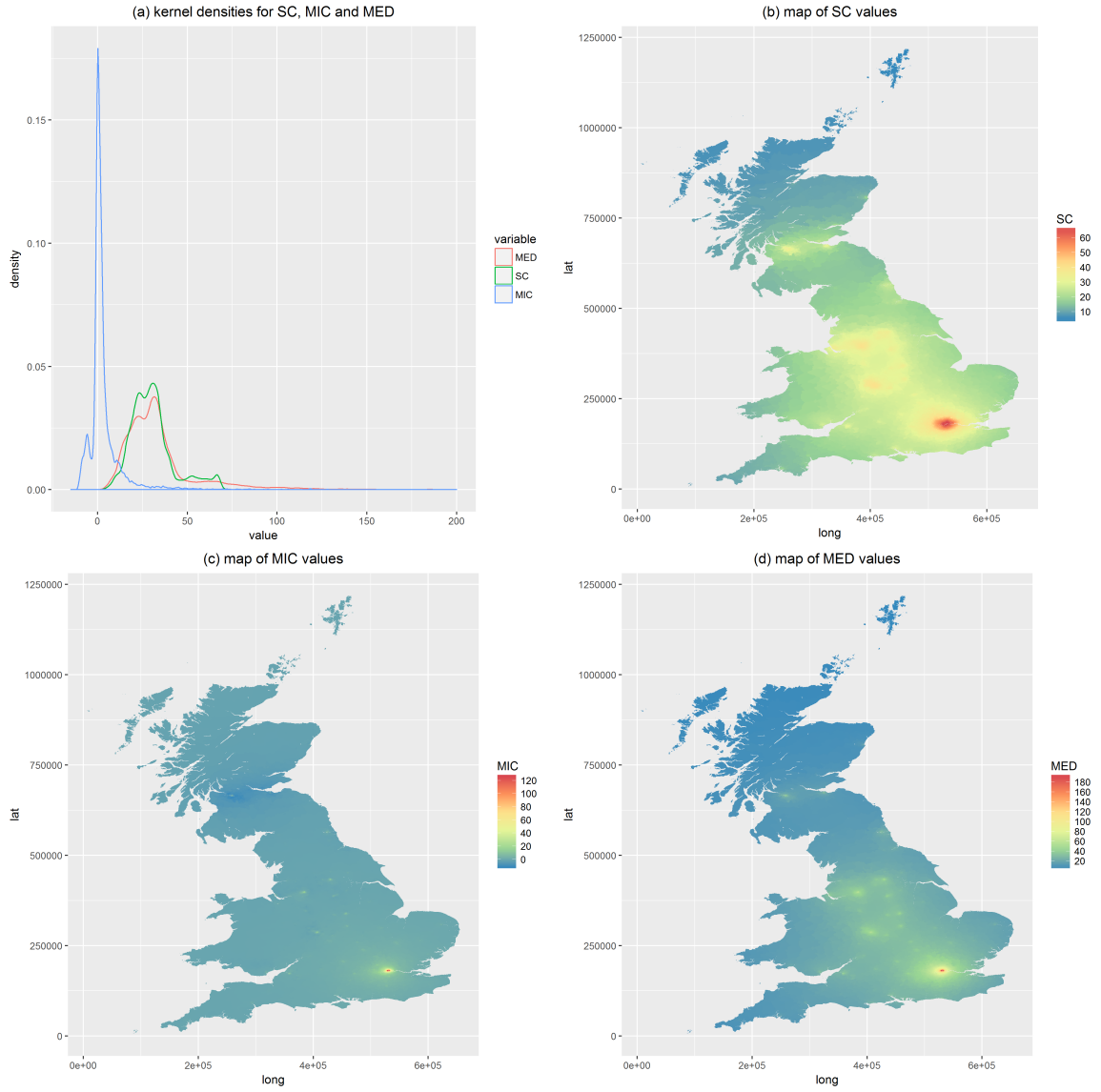

Figure 6: Components of MED values for Britain with total employment as mass

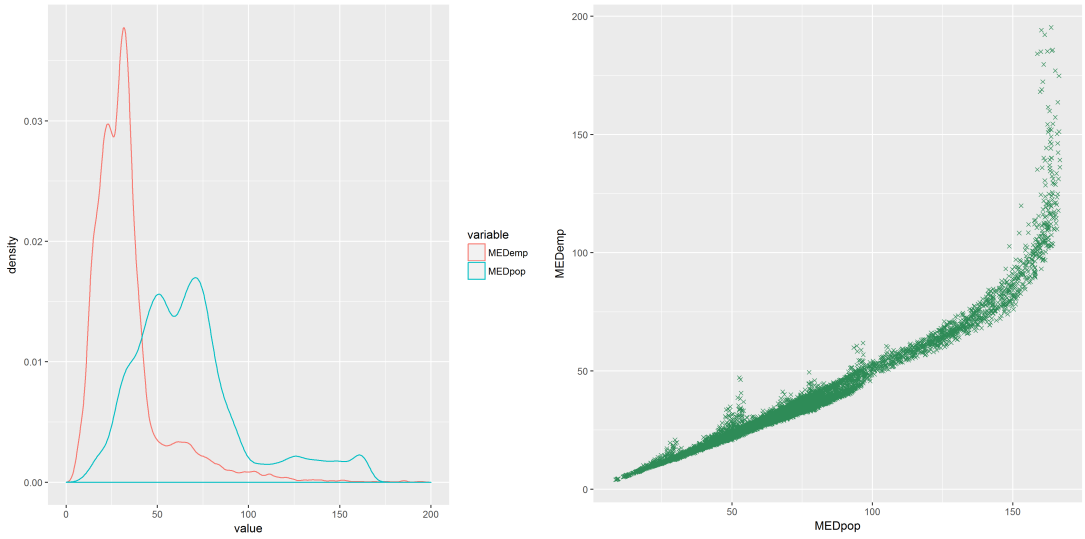

Figure 7: Densities and a scatterplot of MED values for Britain with mass measured by population and employmnet.

4.2 Mass measures

The measures of mass used to construct MED variables provide a means of representing *economic scale* or mass. Possible measures of economic mass at the zone level could include GVA, employment, or population; but the latter two are most commonly used largely due to their being readily availability for small zones over time.

The question then arises as to which mass measure should be used, and also whether different mass measures can offer differ perspectives on agglomeration. For instance, would resident population, rather than workplace based employment, be more appropriate to represent agglomeration effects sourced via labour market connectivity.

In practice, we find that if the mass measures of choice really do provide a good representation of economic scale, then they will tend to produce MED values with similar spatial patterns. This is what we would expect from the decomposition above, given that the MED is highly averaged over space - that is, for each zone, the MED value is a weighted average of mass in all other zones in the study area. The components due to centrality are perfectly correlated with each other regardless of the choice of mass measure. Any difference between the two mass measures is due to the MIC term and if the mass measures follow a similar spatial distribution then these covariances for different mass measures will be very similar too.

Figure 7 shows smooth histograms and a scatterplot for MED values calculated using employment (MEDemp) and population (MEDpop) as mass variables for GB zones. Clearly there are differences between the two measures. In particular, note that for those zones with MEDemp values larger than 100 the slope of the scatterplot becomes very steep. As mentioned previously, these zones are typically those in central London and in the centres of major conurbation which have substantial concentrations of employment. Overall, however, the scatterplot shows that the employment and population EDs provide very similar measures and in fact the correlation between the two is 0.95. It is therefore unlikely that we will learn anything unique about agglomeration effects by using one mass measure rather than the other.

4.3 Impedance functions

The impedance function represents the difficulty experienced in *accessing* economic mass. Candidate measures include distance (Euclidean or route specific), travel time, average speed, the monetary cost of travel, or the generalised cost of travel (e.g. time cost plus monetary costs). For all measures except Euclidean distance, mode specific values could be calculated. In this section we compare some commonly used measures of impedance and comment on their suitability for use in analyses of agglomeration.

There are some basic properties of the formulation of MED measure that are useful for understanding some important aspects of its interpretation. Firstly the scale of measurement for the impedance variable only affects the scale of the index (as does the scale of the mass). An MED index calculated using distances in km will be perfectly correlated with an MED index calculated using distances in miles or cm. This is evident because

$$\frac{1}{n}\sum_{j=1}^{n}\frac{M_j}{(cd_{ij}^{\alpha})} = c^{-\alpha}\frac{1}{n}\sum_{j=1}^{n}\frac{M_j}{d_{ij}^{\alpha}}.$$

A useful implication of this is that when the index is used in a regression analysis in logs (see the section on estimation below), the units used for distance are irrelevant to the estimated agglomeration elasticity (delta). This property also means that conversion of a single network of origin-destination distances into travel time using a fixed travel speed c^1 will generate an MED index that is perfectly correlated with an index using an alternative fixed travel speed c^2. A corollary of this property is that MED indices for different travel modes with different speeds will tend to be very highly correlated, when the origin-destination distances by alternative modes are highly correlated, and also because distances are the main factor in generalised transport costs (see Combes and Lafourcade 2005, for a demonstration for France). We discuss this issue with empirical examples below.

4.3.1 Generalised cost functions

The most common measure of impedance used in transport modelling is generalised cost (GC), which is calculated for distinct modes. The components of GC include the monetary value of time, the fare or price of the trip, and a monetary valuation of trip quality (e.g.. crowding, waiting time penalties etc). GC can vary by time of day for the same trip depending on travel conditions.

As an impedance function for an MED measure, mode specific GC values are appealing because they account for network congestion and thus more accurately measure the true difficulty of accessibility than do distance based measures. Figure 8 shows a scatterplot of MED values for TfL zones calculated using inverse highway GC (ρ^C) and inverse public transport GC (ρ^P) as the impedance functions. The variables are obviously different, as we would expect, but due to the fact that the mass measures are identical for each mode the MED variables they produce similar measures which are highly correlated (0.893).

We expect such correlation to be present across all defined modes and this presents a serious drawback in using mode GC based MED variables for econometric work. Of course, rather than using mode specific MED variables within a single regression function it would be possible to average them in some way into a single measure.

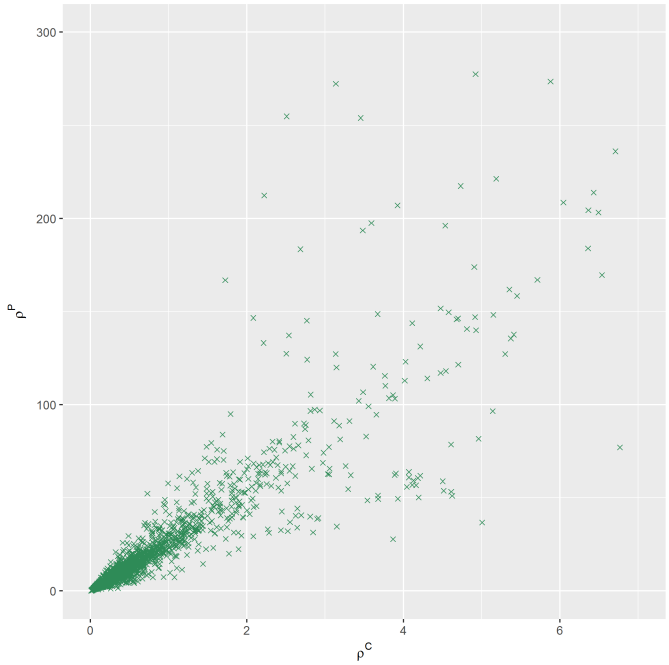

Figure 8: Scatterplot of MED values for TfL zones with impedence measured by highway GC (ρ^C) and public transport GC (ρ^P).

4.3.2 Inverse Euclidean distance functions

Given the potential problem of multicollinearity that arises with high correlation across mode based MED measures, it is appealing instead to have a single measure to represent impedance. This is one of key virtues of using inverse Euclidean distance as an impedance function in MED calculations.

Figure 9 shows that for the TfL data there is a strong association between the mode GC based MED measures and the MED based on Euclidean distance (ρ^D). The correlation coefficient between ρ^C and ρ^D is 0.927, and 0.749 between the ρ^P and ρ^D.

4.3.3 Distance decay

The distance impedance function sometimes includes an exponent on distance (α). Over relatively small spatial scales, such as those of a single city or region, α is often assumed to take a value of 1.0. The empirical literature show that agglomeration effects tend to decrease in magnitude somewhere between 5 and 10 kilometres from source (e.g. DiAddario and Patacchini 2008, Rosenthal and Strange 2008, Melo and Graham 2009). Note that explicit estimation of α for use in transport appraisal can also be undertaken (e.g. Graham et al. 2009). Figure 10 and table 2 show that by increasing α the mean of the MED values falls but the distribution becomes more skewed and the coefficient of variation increases. In effect, when the value of α is increased the influence of the mass of outlying zones on the MED value diminishes in relative terms while that of proximate zones increases.

It is also useful to consider the response of MED to reductions in the impedance, since this is

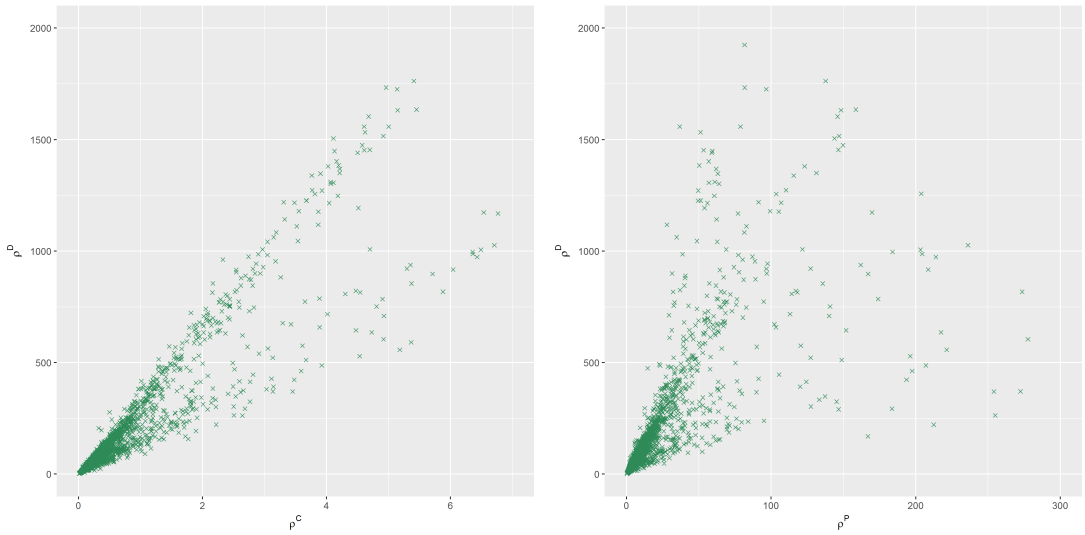

Figure 9: Scatterplots of MED values for TfL zones with impedence measured by highway GC (ρ^C) and public transport GC (ρ^P) against distance (ρ^D).

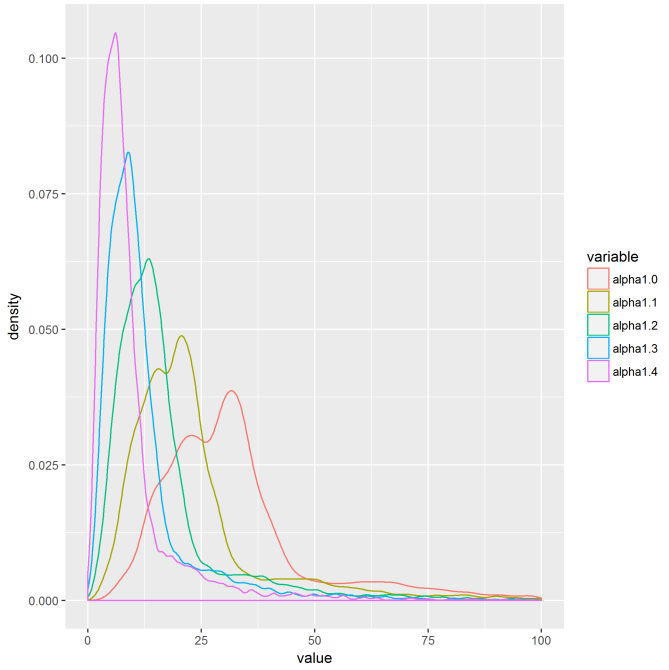

Figure 10: Densities for distance based MED with different distance decay values

Table 2: Summary statistics for MED values for British zones with different distance decay values

α	*1.0*	*1.1*	*1.2*	*1.3*	*1.4*
min	3.947	2.024	1.041	0.536	0.278
median	29.738	19.789	13.456	9.375	6.734
mean	33.647	23.595	17.003	12.605	9.620
max	195.314	176.320	162.294	151.820	143.969
sd	20.973	17.434	14.740	12.678	11.093
sd/mean	0.623	0.739	0.867	1.006	1.153

what transport improvements aim to achieve, and the issue seems to cause a lot of confusion.

Suppose all origin-destination impedances are changed by the same scaling factor, c, due to a general improvement in travel speeds or equivalently a fixed percentage reduction in travel times at all distances. Since $\rho_i = \mathbb{E}(M_j\{cD_{ij}\}^{-\alpha}) = c^{-\alpha}\mathbb{E}(M_j D_{ij}^{-\alpha})$, then

$$\frac{d\rho_i}{dc} = -\alpha c^{-1}\rho_i$$

and the elasticity of ρ_i with respect to c is $\eta_{\rho_i,c} = (d\rho_i/dc)(c/\rho_i) = -\alpha$. Note here that this elasticity of agglomeration with respect to speed is the impedance decay parameter. Other components of the index (original agglomeration level, speeds, distances) are irrelevant. (This is also evident from the discussion of the units of measurement, above).

Now consider a transport improvement which reduces the impedance on one specific link e.g. i_1. Now, using $\rho_i = n^{-1}(\rho_{i1} + \rho_{i2} + \rho_{i3} + \cdots + \rho_{in})$ where, $\rho_{ij} = m_j d_{ij}^{-\alpha}$

$$\frac{d\rho_i}{dd_{i1}} = \frac{-\alpha}{n}\frac{m_1}{d_{i1}^{\alpha-1}} = \frac{-\alpha}{n}\frac{\rho_{i1}}{d_{i1}}.$$

The elasticity $\eta_{\rho_i,d_{i1}} = \partial \log \rho_i / \partial \log d_{i1}$ is $-\alpha/n \times \rho_{i1}/\rho_i$, that is, the impedance decay effect normalised by the relative importance of ρ_{i1} in ρ_i. Thus, a proportional reduction in d_{i1} will always have the absolute effect $\partial\rho_i/\partial \log d_{i1} = -\alpha/n \times \rho_{i1}$. Evaluation at initial d_{i1} and contribution to economic mass provided by destination 1 ($\rho_{i1} = m_1 d_{i1}^{-\alpha}$), will have effect $\partial\rho_i/\partial d_{i1} \times \partial\rho_{i1}/\partial d_{i1} = -\alpha/n$. Note that these are responses to an increase in impedance; a reduction in impedance implies an increase in ρ.

If instead we evaluate the elasticity at the initial mean impedance (d_i) and MED (ρ_i) for origin i (across destinations j) we get $\alpha(\rho_{i1}/\rho_i)(d_i/d_{i1})$. Now the elasticity of effective density w.r.t the impedance reduction on a link i_1 increases (in absolute value) with the effective density at i that is attributable to destination 1 relative to the overall effective density at origin i, and decreases with the distance from i_1 relative to the mean distance from i to other destinations. This is simply telling us that an impedance reduction on a link between an origin and destination will have bigger percentage impact on the MED for the origin i, when that destination was a larger component in the overall MED index for origin i and the destination was closer to it.

Importantly, this does not on its own imply that the impacts on agglomeration are greater when places are generally close together than when they are further apart. To see this, imagine one region A where a number of zones of total mass M are connected by equal impedances d_A and another B where the same number of zones with the same total mass M, are connected by equal impedances $d_B < d_A$. The effective density in A is $\sum(M_j/d_A^{-\alpha} = d_A^{-\alpha}M)$ and the elasticity w.r.t changes in d is just α (evaluated at initial d_A and effective density). The effective density in B is $\sum(M_j/d_B^{-\alpha} = d_B^{-\alpha}M)$ and the elasticity is again just α.

The implication of these results is that the impedance decay parameter (α) is a key parameter determining the sensitivity of the agglomeration index (MED) to changes in impedance. A higher impedance parameter implies that the MED index of agglomeration is more sensitive to reductions in the impedance (e.g. transport cost reductions). Another implication is that transport improvements will have bigger impacts on the effective density of a place when they change connectivity to places that are already important components of its effective density (reductions in travel times to small distant places will not have much impact). The last

implication is that the MED index on its own has no implications for whether its best to target improvements to dense areas or sparse areas, although other related factors the number of firms or workers over which to aggregate the productivity benefits implies by the agglomeration gains, or the initial average productivity of the region, may well have implications when it comes to the final cost benefit analysis.

Use of the exponent α is one way to represent the importance of proximity and the decay of agglomeration with distance. It has the virtue that it requires only a single parameter to be estimated and plugged in to appraisal calculations. There are, however, other mathematical forms and estimation approaches that could be used to represent distance decay (for a review see Graham et al. 2009), including the piecemeal distance band method used to good effect by Rice et al. (2006) and Graham et al. (2009).

4.4 Recommendations on access to economic mass measures

In this section of the report we have shown that MED variables can be measured in different ways using different representations of economic mass and impedance. For the phase 2 empirical work, two key recommendations emerge.

R1 Given the importance of ATEM measures in evaluation of WEIs we recommend that the phase 2 empirical work explore the implications for elasticity estimation of using different measures of mass and impedance for ATEM variables. This will include use of population and employment to represent mass and average GC as well as distance to represent impedance.

R2 Given the importance of the distance decay parameter α in the final agglomeration calculations, we recommend that alternative approaches to estimate distance decay be applied and compared and that implications for appraisal calculations be assessed.

5 Estimating the effect of agglomeration on productivity

The second step in the agglomeration WEI calculation outlined in section 3 uses regression analysis to estimate elasticities of productivity with respect to agglomeration. In this section of the report we explain how productivity can be measured, discuss the main challenges for estimation, and review the econometric methods that can be applied.

5.1 Measuring productivity

As mentioned previously, econometric work on urban agglomeration proceeds by testing for the effect of agglomeration on measures of productivity, typically either Total Factor Productivity (TFP) or labour productivity (LP). TFP is investigated via production function based models, which describe the relationship between the inputs that are used in production and the outputs that are produced. An MED measure of agglomeration can be included within the production function as a shifter of productivity to test for agglomeration effects. A generic production function model form is

$$Y = g(Z, \rho) f(X)$$

where Y is output, X is a vector of input factors, and Z is a vector of factors that affect production levels. The objective is to estimate the agglomeration elasticity, $\partial \log Y / \partial \log \rho$, which will take a positive value if agglomeration economies are present.

Labour productivity is represented by invoking the assumption that workers are paid the value of their marginal product, and by then using the wage rate (i.e. for workers, cities, regions etc) as an implicit measure of LP. Estimating equations for wages can be derived from production functions by assuming optimising behaviour on the part of firms. Alternatively, at the level of individual workers, they can be specified as a so-called Mincerian type wage equation (Mincer 1974) where the wage of each worker in a given location is explained by a set of worker-specific variables (e.g. education, age, gender, skills etc) and a set of 'environmental' characteristics which include agglomeration economies. The generic wage equation model form is

$$W = g(\rho) f(U)$$

where w is the wage rate and U is a vector of covariates relevant for wage determination. Again, the hypothesis that agglomeration externalities exist is tested via the elasticity $\partial \log W / \partial \log \rho$.

5.1.1 TFP versus LP (wage) models

The choice of whether to proceed with a TFP or LP model is often determined largely by data availability or ease of econometric estimation. In relation to assessing WEIs of transport, however, there are some advantages and disadvantage of each approach that we outline here.

First, is the issue of what TFP or wages ultimately depend on in the spatial setting. To address this question Combes and Gobillon (2015) specify a Cobb-Douglas production function for firm i at time t that takes the form

$$Y_{it} = A_{ct} \left(s_{it} L_{it}\right)^{\beta_L} K_{it}^{\beta_K} \tag{2}$$

where c indexes a spatial market area, L is labour input, K is capital, and s is average labour skills.

The profit of the firm is

$$\pi_{it} = p_{it}Y_{it} - w_{it} - r_{it}k$$

where r_{it} is the unit price of capital.

Two important features of this representation are as follows:

1. Agglomeration effects are external to the firm and captured by the shifter A_{ct}, which describes local market 'technology' and may be determined by factors other than agglomeration economies (i.e. $\rho_{ct} \in A_{ct}$). In other words, the estimated 'agglomeration' elasticity can potentially capture non-agglomeration related effects. Subsection 5.2 below reviews econometric methods that can be used to help reduce this source of bias.

2. Labour skills s_{it} are not assumed to be homogeneous across firms, and in fact they could be correlated with local market technology.

Combes and Gobillon (2015) use (2) to derive and compare the assumptions underpinning individual level TFP and wage based approaches to estimation of agglomeration economies. By subtracting the log transformed inputs from log transformed revenue they derive an expression for TFP (ω) which depends on the output price, local technology, and labour skills: i.e.

$$\omega_{j,t} = f(p_{j,t}, A_{c,t}, s_{j,t}).$$

For wages, on the other hand, profit maximization yields the expression:

$$w_{j,t} = f(p_{j,t}, A_{c,t}, r_{c,t}, s_{j,t}).$$

Thus, while TFP is determined solely by local technology, output prices and average labour skills; wages depend in addition on the cost of inputs other than labour, such as land and housing prices. Put simply, the marginal product of any one factor such as labour depends on the quantity of labour relative to other inputs, which in turn depends on the relative prices of inputs. Combes and Gobillon (2015) argue that this renders wage models more susceptible to problems of endogeneity and less easy to interpret.

Gibbons and Overman (2009) pick up this theme in the context of appraising WEIs within CBA. They argue that TFP based measures of productivity should be preferred because they involve direct measurement of returns to scale and productivity; while factor price based measures (e.g. wages, land prices, rents etc) have two key disadvantages:

i. they can be affected by transport improvements via routes other than productivity (i.e. via shifts in labour supply), and

ii. even the productivity effects on factor prices can be ambiguous for the reasons shown by Combes and Gobillon (2015).

A second issue to consider in comparing TFP versus wage based measures of productivity, is the assumed economic behaviour. Both TFP and wage models make assumptions about key

relationships in production, but in the context of productivity analysis the wage based approaches require the more stringent assumption that the wage equal the value of the marginal product in competitive equilibrium. This assumption, as (Combes and Gobillon 2015) note, fails in practice because wages are typically only proportional to Labour productivity, rather than equivalent to it, due to the local monopsony power of the firm.

A third issue is that wage models provide only a partial analysis of productivity in the sense that they consider impacts on one factor alone (i.e. labour). TFP models, on the other hand, look not only at labour productivity but at the productivity of all factors and as such they can be used to isolate the effect of agglomeration on each factor of production as well as on TFP (e.g. Mare and Graham 2009). This can be important because we may expect agglomeration to affect 'technology' in a number of different ways. Thus in relation to the effects of transport investments, which could conceivably affect the productivity of all factor of production, the use of TFP rather than LP may be seems preferable.

These conditions suggest that, whenever possible, productivity should be represented by a direct TFP estimation approach rather than inferred indirectly via factor prices. That said, it is also important to note that both wage and TFP models have been used extensively in the literature to estimate agglomeration elasticities with broadly similar results (e.g. Melo et al. 2009). One advantage of wage-based estimates is that they can be derived from data on workers rather than firms. Datasets on workers are often more easily available than datasets on firms. They also usually contain a wealth of information about workers and their skills which is usually unavailable in firm level data. Controlling for skills when estimating the effects of density (and places more generally) on productivity can be crucially important when trying to distinguish the effects of differences in productivity arising purely from differences in density from those arising from the sorting of workers of different skill levels into different locations. Extending this idea, panel data sets on workers can also be employed to estimate by how much the productivity of individuals changes as they move from low to high density locations, which again provides a way to better control for geographical sorting of workers with different skills. This avenue is not open to estimates based on firm level data, since firms almost never change location intact. In TFP models, sorting is dealt with in a less direct way (that requires stronger theoretical assumptions) by allowing for unobserved heterogeneity in productivity across firms, one component of which may be due to differences in factor quality.

5.2 Econometric methods

Whether TFP or wage models are used to estimate agglomeration elasticities there are methodological challenges that need to be addressed in order to obtain valid causal inference on how productivity responds to a change in the level of agglomeration. There are two fundamental issues at stake. Firstly there is the question of whether an observed correlation between productivity and density represents a causal effect from any area-specific factors associated with density, or whether it is due to sorting of firms of differing inherent productivity into high density and low density places. If sorting is the reason, then improving effective density on its own will have no impact on the productivity of a representative or randomly chosen firm. This issue is the same as the 'people versus place' distinction that surrounds all place-based policy making. Even if the association between productivity and density can be causally linked to place (area-specific factors correlated with density), a second question is whether the place specific factor that is affecting productivity is something that is causally attributable to effective density and something that would be replicated by improving transport connections.

In this section of the paper we discuss these 'endogeneity' issues in more detail, the key estimation challenges and review the econometric methods that are commonly used to address them.

5.2.1 Challenges in estimation

There are various definitions of 'causality' available, but all are fundamentally concerned with understanding what we expect to happen to an outcome Y when policy (or other action) induces a change in D, the 'treatment'. In the context of wider benefits of transport, this means understanding what happens to the productivity of a typical firm or worker when a change in effective density is brought about by improved transport connectivity. The causality framework most commonly used for understanding this concept is the Rubin Causal Model, in which a unit that is subject to an intervention is assumed to have a counterfactual outcome which would have occurred in the absence of the intervention or given some other intervention. Comparison of the average outcome in treated units, or more intensely treated units, and the average counterfactual outcome in those units if untreated, or less intensely treated, would provide the causal estimate of the impact of the treatment. However these counterfactual outcomes for treated units are not observed. The econometric methods used in estimation of the effects of agglomeration on productivity can be understood as ways to try to reconstruct this counterfactual via average outcomes. This implies obtaining estimates of the difference between the average productivity of firms in highly connected places and the average productivity of firms in less connected places, netting out any differences between these two groups of firms that would have given rise to differences in productivity even in the absence of any connectivity differences.

Common spurious influences that can inhibit a causal interpretation of the data are referred to in the literature as sources of *endogeneity*. These arise via the following six mechanisms.

i. Endogeneity via unobserved productivity

The relationship between inputs and outputs is typically imperfectly observed because factors such as input quality, technology, and certain items of capital may not be adequately measured. Furthermore, within the context of a production function, the inputs themselves cannot be treated as truly exogenous because inputs are chosen by the producer in the knowledge of some expected level of productivity (e.g. Griliches and Mairesse 1995, Van Beveren 2012). This implies the existence of a productivity component that is unobserved but important to TFP, and which may be determined in various ways by local technology factors such as agglomeration. If ignored, unobserved productivity can induce bias and inconsistency in estimation of TFP. Furthermore, unobserved productivity is 'transmitted' to factor demand equations via optimising behaviour and thus it affects estimation of factor price models as well as TFP models.

ii. Endogeneity via market selection

According to Ackerberg et al. (2007) the market entry and exit of firms is determined by the comparison of variable profits and sell-off value. A firm will rationally exit the market if its

sell-off value exceeds profit, but both of these are influenced by the unobserved productivity of the firm and its capital stock, and thus market exit decisions are a potential source of endogeneity. In relation to agglomeration, this endogeneity may have a systematic spatial form because we expect firms in locations with high levels of ATEM to be subject to more intense competition, and this could induce less productive firms to exit the market. This has implications for the use of balanced panel data, which by definition, contains only surviving firms and can thus induce a bias in estimation of TFP.

iii. Endogeneity via output price heterogeneity

As noted above, the standard theoretical derivation of TFP and wages lead to expressions that depend on output prices. With degrees of imperfect competition, we expect to find price heterogeneity across firms, but such heterogeneity is typically not observed because output data is usually expressed in monetary units that are closely related to revenue. With revenue based output, firms that exist in local markets with higher prices will have seemingly higher productivity. Again, we expect to observe a systematic spatial tendency in this source of endogeneity since ATEM is a key determinant of spatial competition.

iv. Endogeneity via spatial sorting / functional self-selection

Correlation between unobserved functional or occupational differences and the level of agglomeration can bias the estimated relationship between agglomeration and productivity. In general, this is due to unobserved heterogeneity that arises when firms within the same industry are engaged in different activities across different locations. Most commonly, we observe it when spatial self-selection of labour occurs with high quality workers self-selecting into zones that contain the highest quality jobs. Again, this phenomenon implies unobserved spatial variation in components of productivity. SERC (2009) conceptualise sorting as a *people versus place* distinction, in which the 'place' based effects of agglomeration are obscured by the 'people' based effects of sorting.

v. Reverse causality

The relationship between ATEM and productivity may be simultaneously determined. Higher productivity locations may attract a greater level of private investment over time leading to larger economic mass, and this increase in mass can feedback by raising productivity. Indeed, there is empirical evidence for the existence of bi-directionality (see Graham et al. 2010). Estimation of the production function or wage model should therefore allow for reverse causality between productivity and agglomeration. If it fails to do so estimates of agglomeration economies may be biased and inconsistent.

vi. Endogeneity via confounding / omitted variables

Within our TFP or wage models an MED variable is included to capture the effects of agglomeration externalities. The MED variable, is not however, synonymous with the local technology term A_{ct}, which is instead a composite generic term that allows productivity to

shift due to a number of different potential effects. In other words, the 'pure' effect of agglomeration is just one element of local technology that could affect productivity. Other elements could include specific characteristics of local input and output markets. Since not all of the elements of A_{ct} are observed, or even known, it is likely that the ED variable will capture 'confounding' effects in addition to the marginal effect of agglomeration on productivity. It is important to note that micro level panel data approaches will not address this problem because they adjust for time-invariant confounding at the unit level not the zone or market level. Furthermore, the high level of persistence of the agglomeration variable combined with tendencies for multicollinearity will tend to preclude the inclusion of multiple zone level terms, or zone level individual effects, in a single model,

5.2.2 Econometric methods to estimate agglomeration elasticities

In this subsection, we review the most commonly used econometric approaches for estimation of agglomeration economies with micro level panel data (for an extensive review see Combes and Gobillon 2015). It is important to stress here that the recent empirical literature on agglomeration has developed a strong preference for models based on disaggregate micro-level panel data, over aggregate cross-sectional models, for the following reasons.

1. Micro panel data allow for application of sophisticated methodologies to deal with potential sources of endogeneity.
2. Micro panel model allow dynamics and adjustment in behaviour (i.e. lagged effects) to be studied.
3. The precision of estimation can be increased by using both between unit and within unit variation.
4. The behavioural assumptions inherent in economic theory (i.e. profit maximisation, cost minimisation, competitive equilibrium) have micro foundations and it is thus most appropriate to test theory at a micro level.

Given this preference we limit our review to estimation methods for disaggregate panel data. To set the scene for our discussion of panel models, we define the linear panel model

$$\log y_{ict} = \delta \log \rho_{ct} + f_{ct} + u_{it} + e_{ict}, \tag{3}$$

where f_{ct} is unobserved area effects correlated with ρ_{ct} (e.g. average area skills of workers, due to sorting), u_{it} is unobserved firm effects correlated with ρ_{ct} (e.g. due to sorting of more productive firms into high density places), and e_{ict} is a random error term.

1) Panel fixed effects, within groups or first differenced models

Standard panel data methods adjust for time invariant area, firm or worker effects by differencing data within area, firm or worker units over time.

Suppose we have a repeated cross section or unbalanced panel of firms in zones c, observed in multiple periods t and the area effect f_{ct} in (3) above is fixed over time (for example,

unobservables representing the physical topography of a district). In this case the panel model becomes

$$\log y_{ict} = \delta \log \rho_{ct} + f_c + u_{it} + \varepsilon_{ict}$$

The area component can be eliminated by including area specific dummies or, equivalently, transforming the variables into deviations from their area-specific means

$$(\log y_{ict} - \log \bar{y}_c) = \delta \left(\log \rho_{ct} - \log \bar{\rho}_c\right) + (u_{it} - \bar{u}_c) + (\varepsilon_{ict} - \bar{\varepsilon}_c)$$

This is the within-group panel fixed effects estimator, using areas as the panel groups. Note, if the panel is a balanced panel of firms which do not move over time, and $u_{it} = u_i$ is constant over time (for example, a fixed firm management quality effect), then any correlation of u_{it} with $\log \rho_{ct}$ is also, in effect, eliminated by this transformation. This is the case because $(u_{it} - \bar{u}_c)$ contains only cross sectional variation between firms i within area units c, whereas $(\log \rho_{ct} - \log \bar{\rho}_c)$ contains only time series variation in effective density within area units c. Estimation of δ is based on the changes in effective density over time, within areas.

The fixed-over-time area component f_c and fixed-over-time firm specific components u_i could also be eliminated in a panel of firms using a firm fixed effects estimator, by applying the same transformation in deviations of the variables from firm specific group means.]

$$(\log y_{ict} - \log \bar{y}_i) = \delta \left(\log \rho_{ct} - \log \bar{\rho}_i\right) + (\varepsilon_{ict} - \bar{\varepsilon}_i)$$

Practical applications of panel models like this are limited by a lack of temporal variation in ρ_{ct} in panels which span short periods of time. If ρ is an effective density measure based on economic mass as the numerator and Euclidian distance as the denominator, then any variation in ρ over time in a balanced panel of firms can come only from changes in the economic mass variable i.e. the spatial distribution of employment or population - since distance is constant. Such changes are usually too small to be practically useful.

One case in which fixed effects panel data methods are potentially useful is when estimating wage equations with a panel of workers, some of whom move across areas from one period to the next (e.g. due to job moves).

Now, if the wage equation is

$$\log w_{ict} = \delta \log \rho_{ct} + f_c + \varepsilon_{ict}$$

then a within-individual transformation eliminates the fixed over time worker and area effects.

$$(\log w_{ict} - \log \bar{w}_i) = \delta \left(\log \rho_{ct} - \log \bar{\rho}_i\right) + (\varepsilon_{ict} - \bar{\varepsilon}_i)$$

In this case, delta can be estimated from variation in effective density for individual workers that occurs as they move from one area c to another. This is the approach used in the SERC elasticity estimation discussed below in Section 5.3.2.

An alternative way to eliminate fixed-over-time effects is to difference the data between periods, within the panel group units (rather than transforming to deviations from group means) i.e. for a panel of firms

$$(\log y_{ict} - \log y_{ic,t-s}) = \delta \left(\log \rho_{ct} - \log \rho_{c,t-s}\right) + (\varepsilon_{ict} - \varepsilon_{ic,t-s})$$

Where s is a lag length (e.g. 1 for first differences). This transformation is applied in the dynamic GMM estimators discussed in Section 5.2.2.

Another approach to eliminating fixed-over time unobservables that are correlated with observables is what is sometimes called a 'correlated random effects' estimator. An example of this method is the Mundlak (1978) estimator, which controls for the panel unit unobservables which are correlated with the regressors, by including the group means of the explanatory variables as regressors, rather than differencing them out by first differencing or the applying the within groups transformation.

So, in the above example, this would imply estimating

$$\log y_{ict} = \delta \log \rho_{ct} + \gamma \log \bar{\rho}_c + u_c + \varepsilon_{ict}$$

where $u_c \sim \mathcal{N}(0, \sigma_u^2)$. Although again this is not very helpful if there is little variation in effective density over time, since $\log \rho_{ct}$ and $\log \bar{\rho}$ would be highly collinear.

A key problem with standard panel models is that consistency depends on the absence of time-invariant confounding, which is a strong assumption to maintain in most empirical settings.

2) Panel instrumental variables (IV)

Panel IV approaches are designed to address problems of endogeneity in a dual fashion. Unit level (i.e. firm or worker) individual effects are included in the model to accommodate omitted variable bias from time-invariant confounding; and IV is applied to nullify the effect of other potential sources of endogeneity such as time-varying confounding, measurement error or reverse causality.

The panel IV model can be specified either in levels, as a 'within' panel estimator via fixed effects or correlated random effects; or in difference, with the individual effects differenced out. In either form, the model draws inference from variance within cross-sectional units rather than between them. IVs selected for use in the model must fulfil two criteria: they must be correlated with the endogenous covariates (i.e. relevant) but uncorrelated with the response (i.e. exogenous). For analyses of agglomeration, this implies that valid IVs must be correlated (preferably highly) with agglomeration but uncorrelated with productivity. A common identification strategy proposed in the literature is to use historic lags of population or employment density, or historic transport networks plans, to instrument for current endogenous measures of agglomeration.

In practice, difficulties can arise in using the panel IV approach for estimation of agglomeration effects.

1. **Zone / market level confounding** - the panel IV model forms the within estimator at the unit level (i.e. at the level of the firm or worker), not at the level at which forces of agglomeration operate (i.e. market or zone levels). Consequently, the variance contributing to estimation of the agglomeration parameter is derived via differences in the cross-section as well as over time. Since MED measures of agglomeration tends to be highly persistent, however, typically the majority of identifying variation is derived via cross-sectional differences. As discussed above, this means that bias and inconsistency can arise in estimating the marginal agglomeration effect due to confounding from other elements of local technology.
2. **Invalid IVs** - in practice IVs that are both relevant and exogenous can be hard to find, and in fact when instruments are only weakly correlated with the endogenous regressors,

or when the instruments themselves are correlated with the error term, IV estimation can produce biased and inconsistent estimates. This problem is further confounded by the fact that the available diagnostic statistics do not provide a full-proof means for detecting an inadequate instrument specification.

3) Dynamic panel Generalised Method of Moments (DP GMM)

A similar estimation approach to that of panel IV can be achieved via Generalized Method of Moments (GMM) estimation of a dynamic panel model. This is again an IV estimator, with a linear dynamic panel specification, that allows for a period of adjustment in the effect of agglomeration on productivity and for individual unit level effects that can be either fixed or random. The DP GMM model specifies a dynamic equation (i.e. with lagged response) in both levels and first-differences and uses the time series nature of the data to derive a set of instruments which are assumed correlated with the covariates but orthogonal to the errors. Specifically, lagged first-differences are used as instruments for equations in levels and lag levels as instruments for first-differenced equations. A set of moment conditions can then be defined and solved within a GMM framework to yield consistent estimates of model parameters (for details see Hall 2005).

DP GMM models can be estimated using worker level wage data or firm level production function data. The DP GMM approach is attractive when exogenous IVs are unavailable and / or the dynamics of process under study are of prime interest, it is however subject to exactly the same practical difficulties outlined above in relation to the panel IV model. In addition, it is often argued that the nature of instrumentation inherent in the DP GMM model gives rise to weak IVs that violate the exogeneity condition (see Combes and Gobillon 2015).

4) Panel control function (CF)

Panel CF models, sometimes referred to as structural estimation approaches, offer an alternative to IV models in the context of production function estimation. Under this general approach, structural assumptions concerning firm behaviour are used to derive a proxy for unobserved productivity resulting from endogeneity. Thus, rather than trying to nullify endogeneity via an orthogonal uncorrelated instrument, as in IV estimation, the panel CF approach instead uses two steps to do the opposite: 1) derive a function that is very highly correlated with the endogenous unobserved productivity, and 2) introduce the proxy function into the production function as an additional model component to obtain consistent parameter estimates (for an extensive review of the CF approach see Van Beveren 2012).

There are a variety of different procedures that have been used to derive the function that proxies for unobserved productivity. Olley and Pakes (1996) (OP) used the firm's long run profit maximisation problem to drive an expression for unobserved productivity as a function of investment and capital stock. Levinsohn and Petrin (2003) (LP) criticised the assumption of a monotonic relationships between investment and productivity inherent in OPs model, and instead used the firm's short run profit maximisation problem to derive a proxy function with intermediate inputs and capital as arguments. More recently, Ackerberg et al. (2015) have cited problems of multicollinearity between factor inputs in the OP and LP approaches as a hindrance to identification, and have instead proposed a proxy function based on invertibility of an input demand function with labour choice conditional on the choice of materials.

The CF approach is used to estimate TFP. The effects of agglomeration on TFP can be modelled using either

1. **A one step procedure** - in which an agglomeration covariate is specified within the production function. A one step approach is used when the the calculation of explicit productivity parameters is of key interest.
2. **A two step approach** - in which TFP is estimated from the production function in a first stage model, and the predicted values of TFP then used as the dependent variable in a second stage regression on agglomeration and other spatial variables.

The merit of the one step approach is that it is more efficient, but the two step approach may permit more flexibility in modelling the relationship between agglomeration and productivity including application of non- or semi-parametric causal methods.

A straightforward way to understand the difference between IV and control function approaches is to consider a simple two equation set up

$$y_{it} = \delta\rho_{it} + cx_{it} + u_{it} + e_{it} \tag{E1}$$

$$\rho_{it} = dz_{it} + gx_{it} + hu_{it} + v_{it} \tag{E2}$$

ρ_{it} is explicitly correlated with y_{it} via unobservables u_{it}, whereas z_{it} provides exogenous variation.

In an instrumental variables setting, z_{it} is used to provide exogenous variation in ρ_i that can be used to estimate δ (intuitively since $y_{it} = cbz_{it} + \cdots$, if we can estimate cb and c, then we can estimate b). Instruments can come through theoretical and institutional reasoning on what constitutes a suitable instrument z_{it}, or in the case of GMM derived from an ad-hoc selection of the other variables x_{it} and their lags.

Control function methods involve deriving proxies for u_i from other observable characteristics of i. These might include lagged inputs such as x_{it-1}. In formal control function methods, these observable characterstics and their functional relationship with u_{it} is informed by micro economic theory. However, standard OLS regression can be viewed as a control function method, in which an ad-hoc selection of regressors is inserted in equation (E1) with associated parameters as a proxy for u_{it}.

The CF approach can circumvent the problem of finding valid instruments, but only if the researcher is willing to impose strong theoretical assumptions (in general, control function methods must also involve exogenous variables that determine the endogenous variable but do not determine the outcome directly, or else must rely on non-linearity in functional form).

A recent study of spatial productivity differentials in New Zealand by Mare (2016) has made use of a novel approach to productivity estimation introduced by Grieco et al. (2016), which addresses biases from spatial differentials in both labour quality and input and output prices. Under this approach, it is assumed that firms make optimal choices on labour and intermediate input use, with capital inputs and input prices taken as given. From the first order conditions for firms' profit maximisation it is then possible to estimate firm TFP, infer firm specific input prices, and construct an industry specific index of market power.

In his analysis of New Zealand, Mare (2016) find that heterogeneity in labour quality make a substantial positive contribution to urban / non-urban TFP differentials. However, allowing

for spatial variation in input and output prices, he also finds that typical TFP estimates based on revenue (with assumed homogeneous output prices) and input quantities (rather than expenditure) substantially underestimate the urban productivity premium. In fact, the relative effect of the two is of a similar order of magnitude: adjusting for heterogeneity in labour quality reduces the Auckland productivity premium (relative to other urban areas) from 7.9% to 2.2%; correcting for input and output price differentials raises it once again to 7.9%.

5.3 Econometric evidence on agglomeration for the UK

5.3.1 Agglomeration parameter values used in WebTag

The agglomeration parameter values used for appraisal of UK transport schemes were estimated by Graham et al. (2009). To represent agglomeration they use an ED measure of the form

$$n\rho_i^D = \sum_{j=1}^{n} \frac{m_j}{d_{ij}^{\alpha}}.$$

They use ONS firm level micro panel data, from the Annual Respondents Database (ARD), to estimate TFP within a Cobb-Douglas production function model. They adopt a panel CF approach for estimation to addresses potential sources of endogeneity arising from unobserved productivity, including via heterogeneity in input quality. Agglomeration elasticities are estimated separately for four broad sectors of the economy: manufacturing, construction, consumer services and business services.

A particular innovation of this study is that it allows agglomeration externalities to diminish over distance. It does this by applying non-linear least squares to estimate the distance decay parameter α shown in the ED formula above. The motivation for identifying this parameter is that in assessing the agglomeration benefits of transport investments it is useful to understand the spatial scale over which these externalities are distributed.

The results from this study yield an overall agglomeration elasticity of 0.04 across all sectors of the economy. For manufacturing and consumer services they estimate an elasticity of 0.02, for construction 0.03, and for business services 0.08. The distance decay parameter is found to be approximately 1.0 for manufacturing, but around 1.8 for consumer and business service sectors and 1.6 for construction. This implies that the effects of agglomeration diminish more rapidly with distance from source for service industries than for manufacturing. The relative impact of agglomeration on productivity is, however, larger for services than it is for manufacturing.

The key empirical results of their research are summarised in the table below.

5.3.2 SERC agglomeration elasticities

In research commissioned by the Northern Way the Spatial Economics Research Centre (SERC) at the LSE use micro panel wage data from the Annual Survey of Hours and Earnings (ASHE) to estimate agglomeration elasticities for the UK. The study emphasises the importance of adjusting for heterogeneity in labour quality to obtain a reliable estimate of the pure effect of agglomeration on wages. They discuss this issue in terms of a *people versus*

Table 3: Summary of UK agglomeration elasticity parameters estimated by Graham et al. (2009)

	sic	*agglomeration elasticity*	α
Manufacturing	15-40	0.024	1.122
		(0.002)	(0.127)
Construction	45	0.034	1.562
		(0.003)	(0.159)
Consumer services	50-64	0.024	1.818
		(0.003)	(0.190)
Business services	65-75	0.083	1.746
		(0.007)	(0.144)
Economy (weight aver.)	15-75	0.044	1.659

place distinction, arguing that if skilled workers are attracted to the largest cities then a productivity gradient will be observed due to a 'people' effect, even in absence of 'place' based agglomeration effects. In the wider literature this is referred to as endogeneity via 'sorting' or as a 'people versus place' distinction. To address sorting the authors adjust for people effects by including detailed information on worker characteristics within their wage model as well as workers level fixed effects.

ED measures of agglomeration are used in the SERC study, but calculated for two modes (car and rail) using GC as the impedance factor (SERC 2009). SERC were able to get around the multicollinearity problem, and estimate separate car and rail elasticities in their regressions, by using different levels of spatial aggregation in their ED measures. The rail ED measure was calculated at inter-Local Authority level and the car ED measure at Census ward level. Consequently, the rail measure captures changes in accessibility at an aggregate level and the road measure at a more localised level. The net result is that the elasticities do not actually provide distinct *modal* elasticity values per se, but rather some combined effects of mode and spatial aggregation.

The preferred SERC elasticities, estimated from a model including worker characteristics, industry controls, and fixed effects, are shown below.

Table 4: Agglomeration elasticities estimated by (SERC 2009)

Mode	*Elasticity*
Car	0.069
	(0.016)
Rail	0.049
	(0.014)

5.3.3 Comparing WEbTAG and SERC elasticities

The studies by SERC and Graham et al. (2009) have similar objectives. They both seek to estimate the relationship between ATEM and productivity. Furthermore, both studies recognise that this relationship is obscured by problems of endogeneity, particularity via sorting

/ functional self-selection (i.e. people v place distinction). Each study attempts to correct for this problem. The SERC study does so by explicitly representing labour quality while Graham et al. (2009) use the structural assumptions of the panel CF approach to adjust for unobserved productivity.

While the overall aim of the two studies is similar, they differ in important ways.

1. **Measure of productivity** - SERC uses wages to represent productivity, while Graham et al. (2009) use TFP.
2. **Measure of economic mass** - Graham et al. (2009) use a single distance based ED measure, SERC use ED measures for rail and road with GC as the impedance factor.
3. **Estimation method** - SERC use panel fixed effects in a wage equation with covariates to adjust for worker characteristics. Graham et al. (2009) use a panel CF method within a production function framework.
4. **Industry samples / coverage** - inevitably the coverage of industries, and the representation of economic sectors within the total sample of observations, will differ between the two datasets used in the studies.

Despite these major differences, however, we actually find that qualitatively speaking the Graham et al. (2009) and SERC elasticities are of broadly similar orders of magnitude. This is illustrated in table 5.3.3 below, which shows 95% confidence intervals for key elasticity estimates.

Table 5: Confidence intervals for the Graham et al. (2009) and SERC elasticities

	lower CI	*upper CI*
Graham et al (2009)		
Manufacturing	0.020	0.028
Construction	0.028	0.040
Consumer services	0.018	0.030
Business services	0.069	0.097
SERC		
Car	0.038	0.100
Rail	0.022	0.076

5.4 Recommendations on productivity measurement

In this section we have shown that there are a variety of different ways of representing productivity and of measuring agglomeration-productivity effects. On the basis of the discussion above we make the following recommendations for phase 2 empirical work.

R3 Agglomeration elasticities should be estimated using both wage and TFP models and results compared. Importantly, this should be done using consistent measures of ATEM for the same spatial units over the same time period to ensure comparability of results.

R4 Different econometric models should be applied using different covariate specification to observe the robustness of elasticity estimates to model assumptions. In particular, it will be informative to observe how conditioning on firm, worker and area level controls changes the magnitude of estimated agglomeration effects.

6 Quantifying the agglomeration benefits of transport schemes

The third and final step in the agglomeration WEI calculation outlined in section 3 uses estimated elasticities, and some measure of transport changes, to quantify the WEIs of transport schemes. In this section we discuss alternative ways of making the agglomeration WEI calculation and explain how and in what circumstances they might be appropriate.

6.1 Agglomeration calculations for transport appraisal

Below we describe different ways of calculating agglomeration impacts from transport interventions. In all cases, the quantity being calculated is the total *proportional change in productivity*, i.e. $(\omega^1 - \omega^0)/\omega^0$, that arises from transport induced changes (as reflected in EDs) across different areas ($i = 1, ..., n$) and in some cases different modes ($k = 1, ..., K$). To simplify notation we do not distinguish productivity changes by economic sector, but if a sectoral disaggregation is to be used then the productivity calculations would be made separately for each sector and summed to give the overall effect.

We set the calculations out formally, using log-differentials, i.e. $d\log\omega \approx (\omega^1 - \omega^0)/\omega^0$, since this form shows the key components of each calculation clearly and makes comparison of different calculations straightforward. We take the WebTag calculation as a starting point, with the other calculations being essentially viewed as generalisations of this particular approach. Note, that in the formulas for $d\log\omega$, terms that can be calculated directly from the EDs are placed within square brackets while those outside the square brackets are parameters that must be estimated via an econometric model.

1. **Mixed average modal and distance based calculation (WebATG)**

   The agglomeration calculation recommended in WebTag has an assumed underlying model for TFP (ω) given by

   $$\omega_i = f(\rho_i^D, Z_i),$$

   where ρ_i^D is a distance based MED of the form

   $$\rho_i^D = \frac{1}{n}\sum_{j=1}^{n}\frac{m_j}{d_{ij}^{\alpha}} = \frac{1}{n}\sum_{j=1}^{n}\rho_{ij}^D,$$

   and Z_i represents all other relevant effects on TFP. An econometric model is required to obtain an estimate of the elasticity of productivity with respect to agglomeration. We denote this elasticity by η_{ω,ρ^D}.

   Changes due to transport interventions are measured via an MED based on average GC. Let $\bar{g}_{ij}$ be the average GC on link $i-j$. The corresponding MED is

   $$\rho_i^{\bar{g}} = \frac{1}{n}\sum_{j=1}^{n}\frac{m_j}{\bar{g}_{ij}^{\alpha}} = \frac{1}{n}\sum_{j=1}^{n}\rho_{ij}^{\bar{g}},$$

   with elasticities

   $$\eta_{\rho_i^{\bar{g}},\log\bar{g}_{ij}} = -\frac{\alpha}{n}\frac{\rho_{ij}^{\bar{g}}}{\rho_i^{\bar{g}}} \quad \text{and} \quad \eta_{\rho_i^{\bar{g}},\log m_j} = \frac{1}{n}\frac{\rho_{ij}^{\bar{g}}}{\rho_i^{\bar{g}}}.$$

The productivity calculation applies the elasticity estimated using a distance based MED to changes in MEDs based on average GC. Holding Z_i constant, total proportional change in TFP is

$$d\log\omega = \sum_{i=1}^{n} \frac{\partial \log \omega}{\partial \log \rho_i^D} d\log \rho_i^{\bar{g}} = \sum_{i=1}^{n} \eta_{\omega,\rho^D} d\log \rho_i^{\bar{g}},$$

and total proportional change in the MED can be decomposed as

$$d\log\rho_i^{\bar{g}} = \sum_{j=1}^{n} \frac{\partial \log \rho_i^{\bar{g}}}{\partial \log m_j} d\log m_j + \sum_{j=1}^{n} \frac{\partial \log \rho_i^{\bar{g}}}{\partial \log \bar{g}_{ij}} d\log \bar{g}_{ij}.$$

Under so called 'static' agglomeration calculations the terms of $d\log\rho_i^{\bar{g}}$ involving m_j will disappear because it assumed that $d\log m_j = 0$. Under 'dynamic' calculations all terms $d\log m_j$ and/or $d\log \bar{g}_{ij}$ can change.

Thus, under the WebTag approach, the calculation of proportional change in productivity arising via agglomeration effects takes the form

$$d\log\omega = \sum_{i=1}^{n} \eta_{\omega,\rho^D} \left[\sum_{j=1}^{n} \eta_{\rho_i^{\bar{g}},\log m_j} d\log m_j + \sum_{j=1}^{n} \eta_{\rho_i^{\bar{g}},\log \bar{g}_{ij}} d\log \bar{g}_{ij} \right],$$

where everything within the square brackets can be calculated directly from the data used to construct the GC MEDs.

2. Average modal GC calculation

Rather than apply a distance based elasticity, as in the WebTag approach, we could instead make the calculation using a GC based elasticity. An appropriate underlying assumed productivity model could be $\omega_i = f(\rho_i^{\bar{g}}, A_i, Z_i)$, where A_i captures non transport related agglomeration effects. The corresponding productivity calculation, holding A and Z constant, is

$$d\log\omega = \sum_{i=1}^{n} \eta_{\omega,\rho^{\bar{g}}} \left[\sum_{j=1}^{n} \eta_{\rho_i^{\bar{g}},\log m_j} d\log m_j + \sum_{j=1}^{n} \eta_{\rho_i^{\bar{g}},\log \bar{g}_{ij}} d\log \bar{g}_{ij} \right].$$

where $\eta_{\omega,\rho^{\bar{g}}}$ denotes the elasticity of productivity with respect to average GC effective density.

3. Fully mode specific calculation

The average modal GC calculation could be generalised by disaggregating by mode. For a fully mode specific calculation we could assume the following function for TFP

$$\omega_i = f(\rho_i^1, ..., \rho_i^K, A_i, Z_i)$$

where ρ_i^k, $k = (1, ..., K)$, captures the effects of agglomeration that are generated via transport movements on mode k. The modal agglomeration variables could be represented via MEDs of the form

$$\rho_i^k = \frac{1}{n} \sum_{j=1}^{n} \frac{\theta_{ijk}(g_{ij}) m_j}{g_{ijk}^{\alpha}} = \frac{1}{n} \sum_{j=1}^{n} \rho_{ij}^k$$

where $\theta_{ijk}(g_{ij}) = \theta_{ijk}(g_{ij1}, ..., g_{ijK})$ is the mode share on link $i-j$, specified as a function of the generalised costs (GCs) on that link.

The elasticity of ρ_i^k with respect to the GC of mode k on link $i-j$ is

$$\eta_{\rho_i^k,\log g_{ijk}} = \frac{\partial \log \rho_i^k}{\partial \log g_{ijk}} = \left(\frac{\partial \log \theta_{ijk}(g_{ij})}{\partial \log g_{ijk}} - \alpha \right) \frac{1}{n} \frac{\rho_{ij}^k}{\rho_i^k}$$

and with respect to mass at location j is

$$\eta_{\rho_i^k,\log m_j} = \frac{\partial \log \rho_i^k}{\partial \log m_j} = \frac{1}{n} \frac{\rho_{ij}^k}{\rho_i^k}.$$

Note that the inclusion of mode share in the modal MED variable ensures that the influence of changes in mass or GC on agglomeration are proportionate to the use of that mode. This essentially prevents agglomeration impacts from overlapping across modes.

Holding non transport related TFP effects constant (i.e. A and Z), calculation of proportional change in productivity arising via agglomeration effects could take the form

$$d \log \omega = \sum_{i=1}^{n} \sum_{k=1}^{K} \eta_{\omega,\rho^k} \left[\sum_{j=1}^{n} \eta_{\rho_i^k,\log m_j} d \log m_j + \sum_{j=1}^{n} \eta_{\rho_i^k,\log g_{ijk}} d \log g_{ijk} \right],$$

with parameters η_{ω,ρ^k} estimated from a single productivity regression model.

4. Distinguishing localisation and urbanisation effects

For each of the calculations discussed so far it is possible to generalise further by making a distinction between impacts generated via localisation or urbanisation economies. For sector s, $s = (1, ..., S)$ in zone i, localisation (ρ_i^s) and urbanisation (ρ_i^U) can be represented empirically using employment data E, or some other measure of industry size, by the MED variables

$$\rho_i^s = \frac{1}{n} \sum_{j=1}^{n} E_{sj} f(d_{ij}),$$

$$\rho_i^U = \frac{1}{n} \sum_{j=1}^{n} \sum_{s=1}^{S} E_{sj} f(d_{ij}),$$

where $f(d_{ij})$ could be a function of distance, the GC of a particular mode, or average GC. If an econometric model can provide valid estimates of η_{ω,ρ^s} and η_{ω,ρ^U}, the calculations given above could be expanded accordingly. Note that for a fully mode specific calculation $2 \times K$ parameters would have to be estimated within a single productivity regression model.

5. Calculations with heterogeneous agglomeration effects

In each of the calculations outlined above the productivity elasticities, $\eta_{\omega,\rho}$, are assumed to be constant across zones. We could instead allow for heterogeneous responses of productivity to agglomeration within our econometric model, thus yielding potentially different elasticities for different zones or regions of the country.

6.2 Evaluating alternatives to WebTag calculations for transport appraisal

Having outlined the form of calculations different to that suggested in WebTag, we now provide an evaluation of their potential usefulness in practice.

6.2.1 Average modal GC calculation

As described in calculation 1 above, current UK transport appraisal practice applies agglomeration elasticities estimated using an inverse distance based ED measure to calculate WEIs of agglomeration from transport improvements that are measured by changes in GC. The Department for Transport decided against the use of GC based elasticities on the grounds that it would incorporate an element of double counting with conventional travel time savings (e.g. DfT 2007). The logic here is that since the benefits to business and freight users from congestion reduction are already included in a standard cost benefit analysis, inclusion of congestion effects implicit in GC but not distance ED measures, would risk some element of double counting of benefits.

The empirical work in section 4 above has shown that in practice MED variables will tend to provide similar representations of ATEM for appropriate mass and impedance measures. Furthermore, it is important to note that the conceptual arguments for WEI calculations adopted by DfT, as set out in Venables (2007), effectively view ATEM as synonymous in time and distance. Under this perspective an increase in ATEM due to an increase in economic mass is approximately equivalent to an increase in ATEM due to a decrease in impedance.

In estimating agglomeration elasticities, however, use of GC versus distance based MED variables raises two econometric issues.

1. **Endogeneity** - GC based MED measures will tend to be more endogenous than distance based measures because they incorporate the cost of network congestion, which is in part determined by productivity and by agglomeration.
2. **Complexity** - to construct a single MED GC based measure, rules would have to be introduced on how to incorporate variance across modes and times of day and how to address zero entries in OD matrices when it is not possible to travel from i to j by a given mode.

Aside from these econometric challenges, use of an average GC based MED measure of agglomeration does have one apparent advantage in that, when used in conjunction with a general non-transport related agglomeration term (i.e. A), it may be possible to distinguish between transport and non-transport related agglomeration effects in an econometric model. Use of a distance based MED does not allow such a distinction to be made, and consequently, the WebTag productivity calculation attributes all productivity effects of agglomeration to changes in average GC. In practice, however, it will likely be hard to estimate distinct transport and non-transport related productivity elasticities due to the severe multicollinearity that tends to plague regressions of this sort.

6.2.2 Fully mode specific calculation

Calculation 3 above jointly applies mode specific elasticities to appraise the agglomeration benefits of transport improvements. The application of mode elasticities is appealing because they seemingly provide a more direct match to the types of accessibility changes typically considered in transport appraisals.

We have shown in principal how mode specific calculations of agglomeration effects could be made. In practice, however, it will be hard to obtain the parameters necessary to make such calculations due to econometric difficulties in estimating mode elasticities. Modal MEDs tend to highly correlated potentially giving rise to severe problems of multicollinearity when estimating multiple elasticities $\eta_{\omega,\rho^1}, ..., \eta_{\omega,\rho^K}$ within a single regression model. Since the numerator used to construct the the modal MED variables is identical, a regression model will typically fail to tell us about how each ATEM independently influences productivity.

To illustrate this point, we define two MEDs ρ^C and ρ^R, and note that their inclusion within the same regression model is being akin to estimating

$$\log y = \delta^C \log(W \cdot M) + \delta^R \log(V \cdot M) + \cdots$$

where $\rho^C = W \cdot M$ and $\rho^R = V \cdot M$, with W and V being two different spatial weights matrices (i.e. W for road times and V for rail times). These spatial weights matrices aggregate M in the vicinity of each data point, with weights that decline with the travel time (and hence distance) from each data point to every other data point in the estimation dataset. Crucially, the spatial distribution of M is identical in each case e.g. the spatial distribution of employment in Britain. The only element that differs is the system of weights used to aggregate it.

Such an estimation strategy is surely doomed to fail due to multicollinearity. Indeed, if the only relevant factor differentiating modes is speed, and W and V are constructed for the same zone O-D pairs (as they should be), then $W \cdot M$ and $V \cdot M$ would be perfectly collinear.

Thus we believe it is highly unlikely that the necessary parameters could actually be estimated separately in a manner appropriate to their use in a fully mode specific calculation such as shown above. **For this reason we recommend against attempts to estimate mode based elasticities.**

An additional issue to note in this context is that for any 'dynamic' scenario, in which changes in ATEM are brought about by changes in the location of activity (i.e. economic mass or numerator) in addition to changes in the GC of transport modes (i.e. access or denominator), a joint mode calculation will overlap benefits if MEDs are not normalised by mode share to give a weighted average rather than a sum of the mode-specific elasticities.

6.2.3 Evaluation of urbanisation and localisation economies

As mentioned previously, the theory of agglomeration makes a distinction between agglomeration externalities that are due to industry concentration (localisation economies) and those due to urban concentration (urbanisation economies). We note above that this distinction could be introduced in the productivity calculations as long as we can obtain separate estimates of the elasticity of productivity with respect to urbanisation and localisation respectively.

Previous empirical studies of agglomeration have reported localisation and urbanisation elasticities, but problems of collinearity tend to adversely effect the success of these models. We expect localisation and urbanisation MEDs to be correlated because the denominators are the same and also because we know that industry concentration and urban concentration tend to coincide. Table 6 below shows correlation coefficients between the urbanisation MED and localisation MEDs calculated for SIC sections.

Table 6: Correlation between urbanisation MED and sectoral MEDs

SIC section	*correlation*
A: Agriculture, forestry and fishing	0.752
B: Mining and quarrying	0.094
C: Manufacturing	0.586
D: Electricity, gas etc	0.683
E: Water supply	0.885
F: Construction	0.963
G: Wholesale and retail	0.980
H: Transportation and storage	0.957
I: Accommodation and food	0.991
J: Information and communication	0.970
K: Financial and insurance	0.927
L: Real estate	0.983
M: Professional, scientific and technical	0.977
N: Administrative and support service	0.996
O: Public administration and defence	0.972
P: Education	0.983
Q: Human health and social work	0.979
R: Arts, entertainment and recreation	0.992
S: Other service activities	0.993

The table shows that for most sectors of the economy, correlation between the localisation and urbanisation MEDs is high and is likely to induce problems of multicollinearity. Under these conditions, estimation with a single agglomeration MED variable may be preferred as it will likely capture the combined effect reasonably well.

There are two additional conceptual difficulties with the localisation-urbanisation distinction that arise when applied in the context of transport appraisal. First, is that it is hard to imagine a situation in which a transport intervention alters localisation without simultaneously altering urbanisation. Thus, to treat these two effects as distinct additive components, rather than combining them in a general agglomeration term, may not really add any additional value. Second, it is debatable whether industrial classifications is really the most effective way of defining concentrations of like firms. There is evidence, for instance, that concentrations based on functional characteristics are also prevalent.

6.2.4 Heterogeneous agglomeration effects

The final calculation discussed above notes that agglomeration effects need not be assumed homogeneous across zones. Previous papers by Graham and Van Dender (2011) and Le Nechet et al. (2012) have allowed for nonlinear agglomeration effects and have also estimated separate agglomeration elasticities for sub-samples of the data based on area type or other classifications. This is straightforward to do and can reveal important differences in the responsiveness of productivity to ATEM.

6.3 Recommendations on modal, localisation and heterogeneous agglomeration effects

In this section we have shown that there are a variety of different ways of calculating transport induced productivity effects that arise via agglomeration. On the basis of the discussion above we make the following recommendations for phase 2 empirical work.

R5 Due to limitations of existing data, and econometric challenges arising from severe multicollinearity, we do not recommend that attempts be made to estimate mode specific agglomeration elasticities.

R6 Since the data to represent localisation and urbanisation MEDs are readily available, and since these measures can be easily incorporated into the our main econometric models, we recommend that attempts be made to distinguish localisation and urbanisation effects.

R7 Heterogeneous agglomeration effects should be explored by allowing for non-linearities in the econometric model and by estimating separate agglomeration elasticities for sub-samples of the data based on area type.

7 Recommendations

Below we reproduce our recommendations for phase 2 work as they appear in the text above.

R1 Estimation of agglomeration elasticities based on different measures of mass and impedance for MED variables should be conducted, including use of population and employment to represent mass and average GC as well as distance to represent impedance.

R2 Alternative approaches to estimate the distance decay of agglomeration should be implemented and compared and the implications for appraisal calculations evaluated.

R3 Agglomeration elasticities should be estimated via both wage and TFP models, using consistent measures of ATEM for the same spatial units over the same time period. Results should be compared and a judgement made as to which evidence is most robust and suitable for use in appraisal.

R4 Different econometric models using different covariate specifications should be tested to observe the robustness of elasticity estimates to model assumptions (e.g. conditional versus unconditional estimates).

R5 Due to limitations of existing data, and econometric challenges arising from severe multicollinearity, we do not recommend that attempts be made to estimate mode specific agglomeration elasticities.

R6 Models that distinguish localisation and urbanisation effects should be estimated with a view to deciding whether the resulting evidence is suitable for use in appraisal.

R7 Econometric models should be deigned to explore the existence of heterogeneous agglomeration effects by allowing for nonlinearities and by estimating separate agglomeration elasticities for sub-samples of the data based on area type. Consideration should be given as to whether the resulting evidence is suitable for use in appraisal.

References

Aaberg, Y. (1973). Regional productivity differences in Swedish manufacturing. *Regional and Urban Economics 3*, 131–156.

Ackerberg, D., C. Lanier Benkard, S. Berry, and A. Pakes (2007). Econometric Tools for Analyzing Market Outcomes. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6 of *Handbook of Econometrics*, Chapter 63, pp. 4171–4276. Elsevier.

Ackerberg, D. A., K. Caves, and G. Frazer (2015). Identification properties of recent production function estimators. *Econometrica 83*, 2411–2451.

Ahlfeldt, G. M., S. J. Redding, D. M. Sturm, and N. Wolf (2015). The economics of density: Evidence from the berlin wall. *Econometrica 83*(6), 2127–2189.

Au, C.-C. and J. V. Henderson (2006). Are chinese cities too small? *Review of Economic Studies 73*(3), 549–576.

Baldwin, J., D. Beckstead, W. Brown, and D. Rigby (2007). Urban economies and productivity. Technical Report Economic Analysis Research Paper Series.

Baldwin, J., W. Brown, and D. Rigby (2008). Agglomeration economies: microdata panel estimates from canadian manufacturing. Technical Report Economic Analysis Research Paper Series.

Brulhart, M. and N. Mathys (2008). Sectoral agglomeration economies in a panel of European regions. *Regional Science and Urban Economics 38*, 348–361.

Ciccone, A. (2002). Agglomeration effects in Europe. *European Economic Review 46*, 213–227.

Ciccone, A. and R. E. Hall (1996). Productivity and the density of economic activity. *American Economic Review 86*, 54–70.

Cingano, F. and F. Schivardi (2004). Identifying the sources of local productivity growth. *Journal of the European Economic Association 2*(4), 720–744.

Combes, P. P., G. Duranton, L. Gobillon, and S. Roux (2008). Spatial wage disparities: Sorting matters! *Journal of Urban Economics 63*, 723–742.

Combes, P.-P., G. Duranton, L. Gobillon, and S. Roux (2010). Estimating agglomeration economies with history, geology, and worker effects. In *Agglomeration Economics*, pp. 15–66. National Bureau of Economic Research, Inc.

Combes, P.-P., G. Duranton, L. Gobillon, and S. Roux (2012). Sorting and local wage and skill distributions in France. *Regional Science and Urban Economics 42*, 913 – 930.

Combes, P.-P. and L. Gobillon (2015). The empirics of agglomeration economies. In J. V. H. Gilles Duranton and W. C. Strange (Eds.), *Handbook of Regional and Urban Economics*, Volume 5 of *Handbook of Regional and Urban Economics*, pp. 247 – 348. Elsevier.

Combes, P.-P. and M. Lafourcade (2005). Transport costs: measures, determinants, and regional policy implications for France. *Journal of Economic Geography 5*, 319.

Davis, D. R. and D. E. Weinstein (2003). Market Size, Linkages, and Productivity: A Study Of Japanese Regions. WIDER Working Paper Series 053, World Institute for Development Economic Research (UNU-WIDER).

DfT (2007). *The additionality of Wider Economic Benefits in transport appraisal*. London: HMSO.

DfT (2014). *TAG UNIT A2.1: Wider Impacts*. London: HMSO.

DiAddario, S. and E. Patacchini (2008). Wages and the city: evidence from italy. *Labour Economics 15*, 10401061.

Duranton, G. and D. Puga (2004). *Microfoundations of urban agglomeration economies*, Chapter in Henderson JV and Thisse JF (eds) Handbook of Regional and Urban Economics, Volume 4. Amsterdam: Elsevier.

Fingleton, B. (2003). Increasing returns: Evidence from local wage rates in Great Britain. *Oxford Economic Papers 55*(4), 716.

Fingleton, B. (2006). The new economic geography versus urban economics: An evaluation using local wage rates in Great Britain. *Oxford Economic Papers 58*(3), 501–530.

Gibbons, S. and H. Overman (2009). Productivity in transport evaluation studies. *Working Paper, London School of Economics*.

Graham, D. and K. Van Dender (2011). Estimating the agglomeration benefits of transport investments: some tests for stability. *Transportation 38*, 409–426.

Graham, D. J. (2000). Spatial variation in labour productivity in British manufacturing. *International Review of Applied Economics 14*(3), 323–341.

Graham, D. J. (2005). *Wider economic benefits of transport improvements: link between agglomeration and productivity, Stage 1 Report*. London: DfT.

Graham, D. J. (2007a). Agglomeration, productivity and transport investment. *Journal of Transport Economics and Policy 41*, 1–27.

Graham, D. J. (2007b). Variable returns to agglomeration and the effect of road traffic congestion. *Journal of Urban Economics 62*, 103–120.

Graham, D. J. (2009). Identifying urbanisation and localisation externalities in manufacturing and service industries. *Papers in Regional Science 88*(1), 63–84.

Graham, D. J., S. Gibbons, and R. Martin (2009). *The spatial decay of agglomeration economies*. London: DfT.

Graham, D. J. and H. Y. Kim (2008). An empirical analytical framework for agglomeration economies. *Annals of Regional Science 42*, 267–289.

Graham, D. J., P. Melo, P. Jiwattanakulpaisarn, and R. Noland (2010). Testing for causality between productivity and agglomeration economies. *Journal of Regional Science 50*, 935951.

Grieco, P. L. E., S. Li, and H. Zhang (2016). Production function estimation with unobserved input price dispersion. *International Economic Review 57*, 665–690.

Griliches, Z. and J. Mairesse (1995). *Production functions: the search for identification*, Volume 5067. Boston, NBER.

Hall, A. R. (2005). *Generalized Method of Moments*. Advanced Texts in Econometrics. Oxford: Oxford University Press.

Henderson, J. V. (1974). The sizes and types of cities. *American Economic Review 64*(4), 640–56.

Henderson, J. V. (1986). Efficiency of resource usage and city size. *Journal of Urban Economics 19*, 47–70.

Henderson, J. V. (2003). Marshall's scale economies. *Journal of Urban Economics 53*, 1–28.

Hensher, D. A., T. P. Truong, C. Mulley, and R. Ellison (2012). Assessing the wider economy impacts of transport infrastructure investment with an illustrative application to the north-west rail link project in sydney, australia. *Journal of Transport Geography 24*, 292 – 305.

Holl, A. (2012). Market potential and firm-level productivity in spain. *Journal of Economic Geography 12*(6), 1191.

Kanemoto, Y., T. Ohkawara, and T. Suzuki (1996). Agglomeration economies and a test for optimal city sizes in japan. *Journal of the Japanese and International Economies 10*, 379 – 398.

Lall, S. V., Z. Shalizi, and U. Deichmann (2004). Agglomeration economies and productivity in indian industry. *Journal of Development Economics 73*, 643 – 673.

Le Nechet, F., P. Melo, and D. Graham (2012). Transportation-induced agglomeration effects and productivity of firms in megacity region of paris basin. *Transportation Research Record*, 21–30.

Levinsohn, J. and A. Petrin (2003). Estimating production functions using inputs to control for unobservables. *Review of Economic Studies 70*, 317–341.

Mackie, P., D. J. Graham, and D. Laird (2012). *Direct and wider economic benefits in transport appraisal*, Chapter in A de Palma, R Lindsey, E Quinet, and R Vickerman (eds) Handbook in Transport Economics, pp. 501–526. London: Edward Elgar.

Mare, D. (2016). Urban productivity estimation with heterogeneous prices and labour. Working papers, Motu Economic and Public Policy Research.

Mare, D. C. and D. J. Graham (2009). *Agglomeration elasticities for New Zeland*. Auckland, Land Transport New Zealand.

Mare, D. C. and D. J. Graham (2013). Agglomeration elasticities and firm heterogeneity. *Journal of Urban Economics 75*, 44–56.

Marrocu, E., R. Paci, and S. Usai (2013). Productivity growth in the old and new europe: The role of agglomeration externalities. *Journal of Regional Science 53*, 418–442.

Martin, P., T. Mayer, and F. Mayneris (2011). Spatial concentration and plant-level productivity in France. *Journal of Urban Economics 69*, 182–195.

Melo, P. and D. J. Graham (2009). Agglomeration economies and labour productivity: evidence from longitudinal worker data for gb's travel-to-work areas. *SERC Discussion Papers* (No SERCDP0031). http://www.spatialeconomics.ac.uk/textonly/SERC/publications/download/sercdp0031.pdf.

Melo, P., D. J. Graham, and R. B. Noland (2009). A meta-analysis of estimates of urban agglomeration economies. *Regional Science and Urban Economics 39*, 332–342.

Mincer, J. A. (1974). *Schooling, experience, and earnings*. New York, Columbia University Press.

Mion, G. and P. Naticchioni (2005). Urbanization externalities, market potential and spatial sorting of skills and firms. Technical report, CEPR Discussion Papers 5172. Centre for Economic Performance, London School of Economics and Political Science.

Moomaw, R. L. (1981). Productivity and city size: a review of the evidence. *Quarterly Journal of Economics 96*, 675–688.

Moomaw, R. L. (1983). Spatial productivity variations in manufacturing: a critical survey of cross sectional analyses. *International Regional Science Review 8*, 1–22.

Moomaw, R. L. (1985). Firm location and city size: reduced productivity advantages as a factor in the decline of manufacturing in urban areas. *Journal of Urban Economics 17*, 73–89.

Morikawa, M. (2011). Economies of density and productivity in service industries: An analysis of personal service industries based on establishment-level data. *The Review of Economics and Statistics 93*(1), 179–192.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica 46*(1), 69–85.

Nakamura, R. (1985). Agglomeration economies in urban manufacturing industries: a case of Japanese cities. *Journal of Urban Economics 17*, 108–124.

Olley, G. S. and A. Pakes (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica 64*, 1263–1297.

Rice, P., A. J. Venables, and E. Patacchini (2006). Spatial determinants of productivity: analysis for the regions of Great Britain. *Regional Science and Urban Economics 36*, 727–752.

Rosenthal, S. and W. Strange (2008). The attenuation of human capital spillovers: a manhattan skyline approach. *Journal of Urban Economics 64*, 373–389.

SERC (2009). *Strengthening economic linkages between Leeds and Manchester: feasibility and implications: full report*. London: SERC.

Sveikauskas, L. (1975). The productivity of cities. *Quarterly Journal of Economics 89*, 392–413.

Sveikauskas, L., J. Gowdy, and M. Funk (1988). Urban productivity: city size or industry size. *Journal of Regional Science 28*, 185–202.

Tabuchi, T. (1986). Urban agglomeration, capital augmenting technology, and labour market equilibrium. *Journal of Urban Economics 20*, 211–228.

Van Beveren, I. (2012). Total factor productivity estimation: A practical review. *Journal of Economic Surveys 26*, 98–128.

Venables, A. J. (2007). Evaluating urban transport improvements: cost-benefit analysis in the presence of agglomeration and income taxation. *Journal of Transport Economics and Policy 41*(2), 173–188.

Venables, A. J., J. Laird, and H. Overman (2014). *Transport investment and economic performance: Implications for project appraisal*. London, DfT.

Wheeler, C. H. (2001). Search, Sorting, and Urban Agglomeration. *Journal of Labor Economics 19*(4), 879–899.