Optimal infrastructure reinvestment and supply in urban rail systems: A dynamic optimization approach

Praj Xuto[1], Richard J. Anderson, Daniel J. Graham, Daniel Hörcher

Railway and Transport Strategy Centre (RTSC)
Department of Civil and Environmental Engineering
Imperial College London

Paper submitted for presentation at the
Annual Conference of hEART, Budapest, 2019

Background

The state of infrastructure in many developed countries around the world is currently an oft-discussed topic, with politicians, citizens, engineers and academics all weighing in with their ideas. It is an increasingly pressing issue, with ballooning costs the longer repairs are deferred. Indeed, in the US, the American Society of Civil Engineers estimated in 2017 that it will cost nearly $4.6 trillion to fix the country's infrastructure to a 'state-of-good-repair' and meet forecast demand by 2025. Perhaps one of the most important aspects of infrastructure is transportation, as it affects millions of people on a daily basis, in a very visible way, and is a necessity for most forms of work and commerce, from the daily commute to freight transport. In today's increasingly urbanizing world, one transport mode in particular stands out – the urban rail network. Indeed, this is especially true in the 21st century where the most important world cities such as London, New York, Tokyo, etc. (Globalization and World Cities (GaWC) Research Network, 2016) generate an outsized share of not only the economic output, but also wield cultural and political influence within their countries, so it is natural to pay particular attention to the infrastructure in them. Moreover, rail is one of the most crucial transport components of a city, as it can provide large volumes of movement at relatively high speeds, enabling workers to get to their workplaces at relatively low economic costs, compared to other modes in large cities.

Focusing more closely on the rail transport organisations themselves, in most systems, the two departments with the largest budgets are typically Operations & Maintenance (O&M) and Infrastructure Construction due to the large number of personnel for the former and construction needs of the latter. The two are directly intertwined because ultimately, the service operations department and passengers will be the users of the new infrastructure. Thus typically, capital renewal spending (capex) on good quality infrastructure that is easy to maintain ought to lower future O&M costs while at the same time, low quality maintenance (or increased service operations) ought to increase future capex needs, all else being equal. Ideally, a balance needs to be struck between the two to optimize lifecycle costs, so analysing ways to optimize the budget breakdown is important; indeed, current asset management research

[1] Proposed presenter – Email address: p.xuto17@imperial.ac.uk

focuses heavily on this, typically at a programme-level of similar projects. On the other hand, the transport economics supply and pricing literature largely overlooks this critical relationship, perhaps due to the early focus on buses (e.g. Mohring, 1972), where capital costs are considerably less than for rail. In fact, as research by the Railway and Transport Strategy Centre (RTSC) & World Bank (2018) demonstrates, on average, older metros spend the equivalent of 37% of their total O&M costs just on reinvestment in the existing system alone (this excludes any expansion investment), a significant amount. Thus, this paper aims to include this key element of rail system costs in order to examine the interplay between asset condition, supply and pricing over time – tying together the asset management and transport supply-pricing literature more definitively.

Related Literature

In the mainstream literature of public transport economics, infrastructure- and asset management-related modelling efforts are normally limited to *investment appraisal*, i.e. the evaluation of the efficiency of building or buying new assets (Small & Verhoef, 2007). Meanwhile, in the case of short-run policy optimisation, the costs of service provision is usually limited to operational expenses directly linked to the capacity (frequency and vehicle size) supplied. On the other hand, in the (relatively new) asset management literature, the reverse is true whereby the focus is the optimisation of asset lifecycle costs through budgeting or work scheduling, with little emphasis on the links to transport demand and short-run policies applied by the operator. For further details of transport asset management, see Sinha et al.'s (2017) review of the field.

From the transport welfare economics perspective, the lack of research is because of the nature of short-run welfare maximisation (as well as perhaps the initial focus on bus networks, as mentioned earlier), implying that all infrastructure costs must be automatically covered by the government since the focus is on marginal costs only. This is perfectly valid but at the same time presents a contrast with private firms, where the profit-maximizing goal is reached when marginal revenue (price) is equal to marginal cost in the short run – yet there is a shutdown condition in the long-run where price must also be greater than average variable cost (in effect, average total cost per unit output since all costs are variable in the long-run). This condition is absent in transport welfare analyses, meaning that infrastructure costs are somehow never considered in the pricing, even though without infrastructure, rail systems would not be able to operate in the first place.

In addition, leaving aside the costs involved with building new systems, it could be argued that non-expansion capital investment (reinvestment) should be considered as part of the optimisation process, since these are recurring costs needed just to keep the asset condition in a steady state, similar to the way maintenance costs are included in analyses. Of course, reinvestment is 'lumpy', but these costs could simply be distributed over the asset's life. Interestingly, Mohring (1972) did also include bus capital cost (cost of the buses themselves) into his short-run optimisation of peak demand, to capture costs of additional service. Bearing these points in mind, an argument could hence be made that i) renewals are essentially just another form of maintenance and/or ii) increased renewals would likely be needed if service levels increases, so can be considered as a form of marginal cost. After all, regular rail infrastructure reinvestments are critical for the ongoing operations of trains.

Another important issue with short-run welfare maximisation is that it ignores the considerable funding challenges that transport companies face in reality. Due to the very high capital requirements even just for reinvestment, some large urban rail systems struggle to obtain enough capital subsidies. The problem is compounded by the fact that renewal spending is often behind-the-scenes, which further reduces the political incentive of the government to fund it. A case in point is the Washington Metro, which has experienced increased infrastructure-related 'incidents' and derailments over the last 10 years, even as the new Silver Line was being built. Additionally, obtaining public funds through taxation is costly and generates deadweight loss on society, as research increasingly shows (e.g. Sun et al., 2016; Proost & Van Dender, 2008). This deadweight loss is typically estimated at around 1.2, meaning a loss of 0.2 for every 1 monetary unit raised (Small & Verhoef, 2007, p. 178), though estimates can vary considerably. As noted earlier, since the short-run model assumes that all infrastructure costs are subsidised, coupled with any operating subsidies (as often suggested by the literature, e.g. Parry & Small, 2009), large sums of money must be generated through taxes, creating considerable welfare losses even if the government is able to fund the rail system adequately.

This theme of further refining the model has been explored in various ways in more recent times, for example, through the inclusion of the cost of public funds as mentioned earlier. Another important consideration is crowding costs (e.g. Tirachini et al., 2013; Hörcher et al., 2017), which is a reality in any busy system, e.g. passengers may choose an alternate route, even if it was longer, to avoid overcrowded sections. Including this cost could further reduce the benefits of low fares (i.e. high subsidies) since many rail systems in major cities are already at or above capacity during the peak, so any additional passengers would impose an external cost on others.

Towards this same goal of model improvement, this work combines an analysis which includes asset-related spending and the impacts of condition on user and operator costs, thereby providing a new focus on the long-term. Thus, the research seeks to directly combine elements from both the economics and asset management literature to provide a more holistic insight for urban rail systems.

Research Objectives

The goal of this work is to develop an economic model of urban rail transport supply in which (i) *infrastructure reinvestment* is an explicit decision variable of the supplier, and (ii) *asset condition* is an intermediate output of production with an impact on both operational costs and demand through the quality of service. Using this model, one can derive the optimal level of reinvestment, the resulting asset condition, and optimal fare levels and subsidies.

Methodology

This work utilizes a dynamic optimization approach, which is also relatively rare in transport pricing analyses, to investigate the relationship between the four decision variables of: fare (P_t), transport supply (train-kms, Y_t), maintenance spending M_t, and renewal spending R_t, where t is the index of time periods. Specifically, the transport organisation optimisation model is formatted as a dynamic programming problem where the objective function is subjected to an intertemporal constraint that is a state variable, in this case asset condition. Essentially, the optimisation in each period is also dependent on the asset condition A_t ('stock') at the start of the period. A simulation framework is implemented due to the simultaneous dependency between multiple production and demand related model components. The overall model focus is at the aggregate firm-level, with each time period equal to one year.

To briefly illustrate the workings of the model, an early output of the simulation experiment is shown below in Figure 1. This graph plots the welfare maximising asset condition as a function of the number of time periods considered in the dynamic optimization (in this case, 20 years). Note that the initial asset quality in period 1 is set to 0.8 exogenously, to reflect current asset conditions.

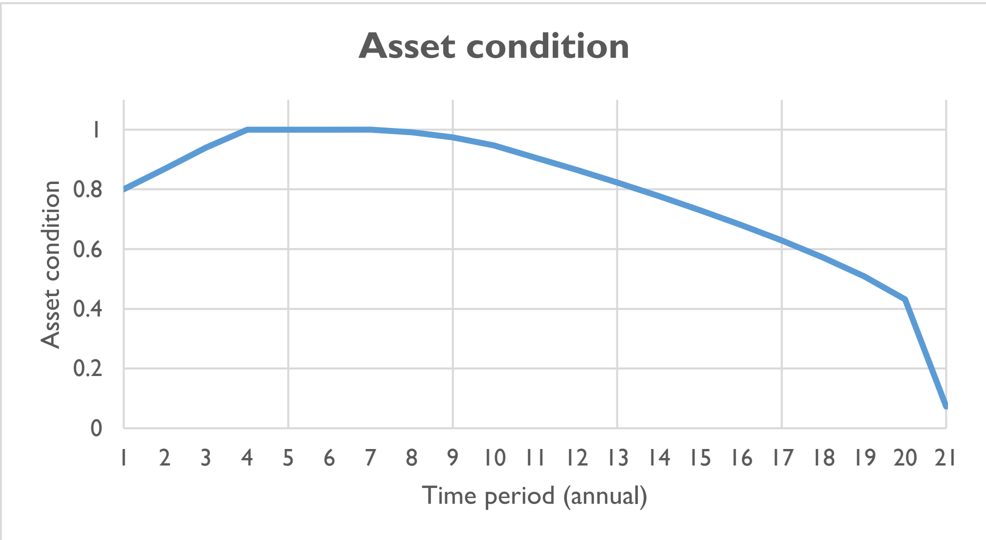

Figure 1

Notice that the asset condition is improved until it reaches a steady state at 1, before gradually starting to fall from around period 7. This is because the goal is welfare maximisation within a set time horizon, meaning that there is a tendency to let the condition decline by spending less on R_t and M_t (see Figure 3 on p. 9) as t approaches n, thereby boosting the welfare. This effect is more clearly seen if n is set to a higher value. (Note that there is also a constraint in period 21 where $A_t > 0$ otherwise the asset condition would have dropped sooner). This gradual decline of the 'stock' variable is similar to the standard example of a dynamic programming consumption model, where the capital stock in the last period is zero since there is no incentive to save for the future. By the recursive nature of the Bellman equation, the stock is gradually used up as t approaches n. Further discussions of the preliminary results and the links between A_t and the decision variables can be found later in this paper.

Returning to the analytical model that is the basis for the simulation, the initial welfare-maximisation objective function is shown in equation (1), with the asset condition intertemporal constraint in equation (2). A_t is meant to represent the overall infrastructure condition of the rail system as a value between 0 and 1, with one equivalent to brand new condition and zero to completely broken assets.

$$max\ W = max \sum_{0}^{n} \int_{0}^{Q_t} D(Q_t)\, dQ_t - C_{agency,t} - C_{users,t} \quad (1)$$

$$s.t.\ \ A_{t+1} = A_t - \frac{1}{j^{A_t}} M_t^k - lY_t + hR_t \quad (2)$$

Where W is the total social welfare, and the maximisation objective is based on net benefits over n time periods, comprised of: $D(Q_t)$ the inverse demand as a function of quantity demanded Q_t (number of passengers); $C_{agency,t}$ the total transport agency costs; and $C_{users,t}$ the total generalised (non-monetary) user costs. These are all values at a given time period t.

For the intertemporal constraint, the asset condition at the start of period $t+1$, A_{t+1}, is dependent on the asset condition at the start of period t, adjusted by the impacts from maintenance spending M_t, transport supply Y_t and renewal spending R_t. The asset condition is assumed to experience exponential decay, where $j > 1$, to reflect the fact that rail infrastructure is a combination of structural (e.g. tunnels, rails) and equipment (largely rolling stock) components. Based on Coen (1975)'s examination of the manufacturing industry, *capacity* depreciation for structures tends to follow a 'one-hoss shay' pattern whereby there is minimal to no output decline during the course of the asset's life but with a sudden catastrophic failure at the end, while equipment depreciation tends to follow either a linear or geometric decay pattern. Combining these two together, an exponential function can best mimic the overall behaviour where there is some limited decay at first which sharply increases near the end of life. The maintenance spending is a counteracting force to this condition decline, thus k is a negative number; it is also inelastic to reflect the fact that additional M_t will have less of an impact on asset condition. Furthermore, Y_t has a negative linear impact on condition due to wear and tear, while R_t has a positive linear impact (both l and h are positive).

The transport agency costs comprise of service operations O_t, maintenance M_t and renewal costs R_t, as shown in equation (3). In addition, there is marginal cost of public funds ϕ if subsidies S_t ($S_t = C_{agency,t} - P_tQ_t$) are needed to cover the total agency cost, however note that if $S_t < 0$, ϕ drops to zero.

$$W = max \sum_{0}^{n} \int_{0}^{Q_t} D(Q_t)\, dQ_t - O_t - M_t - R_t - \phi S_t - C_{users,t} \quad (3)$$

Notice that maintenance is separated out as its own decision variable (in most models, it would typically be included as part of the operating cost), while renewal is now an explicit decision variable, which would not have been considered in static models.

The user costs (equation (4)), in turn, are based on typical journey time costs of access, waiting and in-vehicle travel times. Moreover, crowding and asset condition are factors affecting in-vehicle travel time; both these are relatively recent additions to the 'standard' generalised cost model, particularly the latter.

$$C_{users,t} = v_{acc} T_{acc} Q_t + \frac{v_{wait} Q_t}{2F_t} + \left(v_{inveh} T_{inveh} + v_{inveh} T_{inveh} \psi \frac{\gamma Q_t}{Z Y_t} \right) Q_t A_t^r \quad (4)$$

Where v represents the monetary per hour value of access, waiting and in-vehicle travel times, as appropriate; T the time taken for each type of travel; F_t the system-wide average route frequency per hour in each direction, weighted by route length; ψ the crowding multiplier; γ the average distance travelled per journey; Z the average total floor area (both standing and seating) per train; and A_t^r the asset condition multiplier where $-1 < r < 0$, which is similar to the earlier capacity depreciation logic in that as A_t decreases, the multiplier increases slowly at first before rapidly increasing as condition approaches zero. This is also calibrated with empirical research (Spy Pond Partners et al., 2018) that estimated a multiplier of 1.2 on travel time if the asset is in poor condition.

At equilibrium, $D_t = c_t + P_t$, where c_t is the average user cost and P_t the fare. Using the relationship $Y_t = 2F_t su$ (s is the total system length, u total hours of system operations; the "2" is to capture both directions) and rearranging, Q_t can be written as:

$$Q_t = \frac{v_{acc} T_{acc} + \frac{su v_{wait}}{Y_t} + v_{inveh} T_{inveh} A_t^r + P_t - b}{a - v_{inveh} T_{inveh} \psi \frac{\gamma}{Z Y_t} A_t^r} \quad (5)$$

$$Q_t \equiv z(Y_t, A_t, P_t)$$

The service operations cost is also linked to the decision variables through the relationship in equation (6), so that both the level of supply and asset condition has an impact on cost:

$$O_t = J A_t^{\alpha} Y_t^{\beta} \quad (6)$$

where J is the productivity factor, and α and β the elasticities. The sum of α and β is assumed to be equal to or less than one, i.e. constant or decreasing returns to scale, where $-1 < \alpha < 0$

but is likely to be closer to zero due to the nature of A_t^{α} in this model (if α was closer to -1, the rate of increase of the service operations cost would be unrealistically rapid as condition falls), while β would be around 1 since the relationship between output and O_t is likely to be linear.

This initial model is then simulated through testing multiple objective functions and scenarios behind asset management and service provision, with varying time horizons and goals. Such objectives include social welfare maximisation, profit maximisation, and alternative objectives related to the political economy of asset management. Data for calibration is readily available from the global metro benchmarking groups facilitated by the Railway and Transport Strategy Centre at Imperial College London.

Finally, key model assumptions include:

1. A relatively large, older, steady-state urban rail system is considered – this is to increase the level of aggregation and assume changes are more gradual (as opposed to a smaller system with lower mode share where ridership can fluctuate more). In addition, the large asset base helps to smooth out the natural 'lumpiness' of capex.
2. Only infrastructure (capital) spending on renewals is considered. Spending on system expansion, such as new lines or stations, is excluded.
3. Capacity is assumed to be unbounded, similar to the transport economics literature, however there is crowding cost and the rail system core is assumed to be overcrowded during the peak in the initial state, similar to most major large, dense cities around the world.
4. A dynamic temporal model is used (multiple time periods), which is quite different from the static model used in the transport economics literature. This is because depreciation, capex and condition have temporal dimensions.
5. In order to reflect the real-world differences between maintenance and capex, an important distinction between the two is made, namely:
    a. Capex has a significantly greater impact at increasing the asset condition than maintenance.
    b. There is a 'natural' asset deterioration rate so that the condition will still decline over time given a constant maintenance spending level (the second term on the right hand side of equation (2)).

The model is initially calibrated with London Underground's 2016 data from the RTSC. Another key assumption is that the current levels of P_t, Y_t, M_t, R_t (in real terms) will maintain the same ridership and asset condition into the foreseeable future, so that the system is in a steady-state but is not necessarily at the optimal point.

Expected & Preliminary Results

Numerical results are currently in development, and are expected to provide insights into the interplay between pricing and asset reinvestment. The optimal values of transport supply, fare, maintenance spending and renewal spending will be the key outputs, which could vary depending on the objective function. The data will be plotted over time to demonstrate the effects of different objectives/scenarios (policies), hence providing a long-term perspective of an urban rail system at equilibrium.

Some preliminary results are shown below, with Table 1 an example of testing various scenarios with changing parameters:

Scenarios	Maximum welfare (in millions, rounded to nearest hundred)	% change from scenario 0a:
Scenario 0: Status quo	£107,900	n/a
Scenario 0a: Optimised welfare from status quo	£135,100	0
Scenario 1: Optimised, no crowding cost	£148,700	10.1

Table 1

In addition, graphs of the optimisation from the status quo (Scenario 0a in Table 1) are provided, showing the temporal profiles of key results and variables, with each data point a value in that year. This is shown in Figures 2-5, as well as Figure 1 ("Asset condition") on p. 4. The annual welfare value (Figure 2) is also driven by a similar dynamic to asset condition, where the initial low levels are due to high R_t to improve condition; a steady-state of welfare is reached between periods 4-6, which is then further boosted due to reduced spending before peaking in period 10, and then gradually declining as decreasing A_t takes its toll. Note here that the initial level of R_t (Figure 3) is affected by the constraint, which in this case is £4,000m and is already quite high, considering the current renewals spending of London Underground is in the region of £1-£1.5 billion. In Figures 4 and 5, supply and price, respectively, also work in a similar way; the reason for the former's increase and the latter's decrease, bearing in mind there is cost of public funds, is that the asset-related spending was decreased instead.

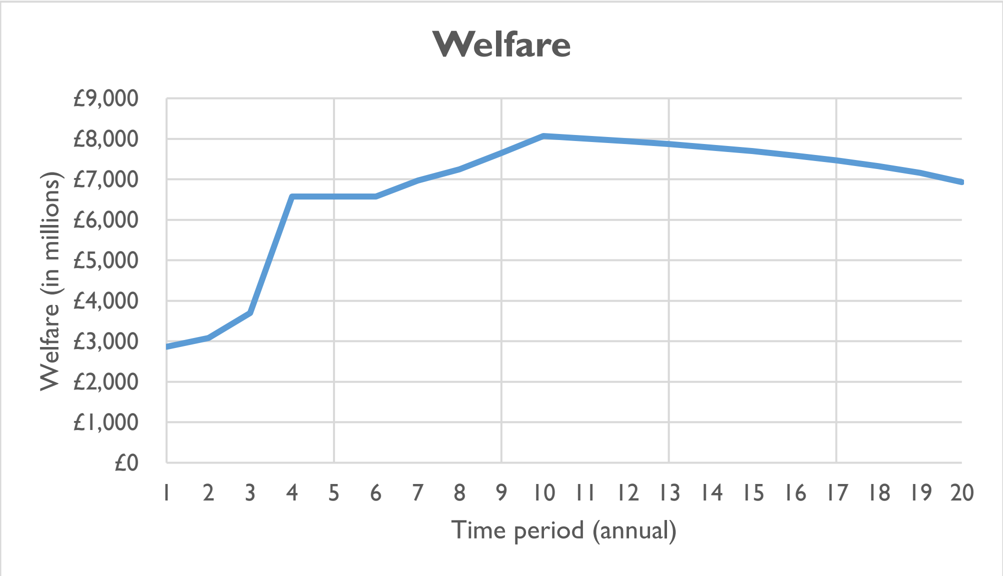

Figure 2

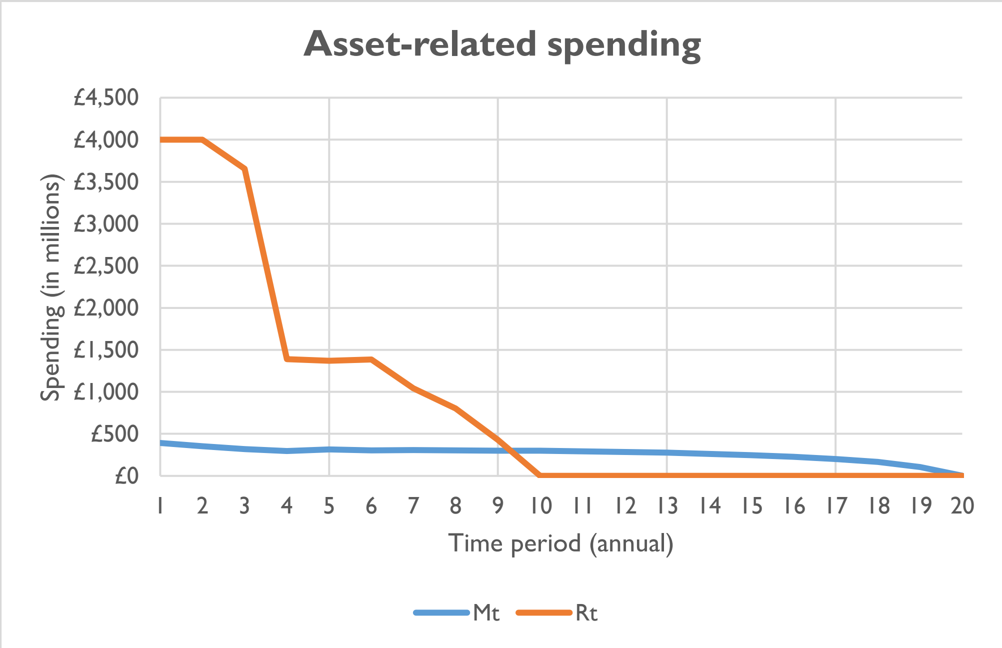

Figure 3

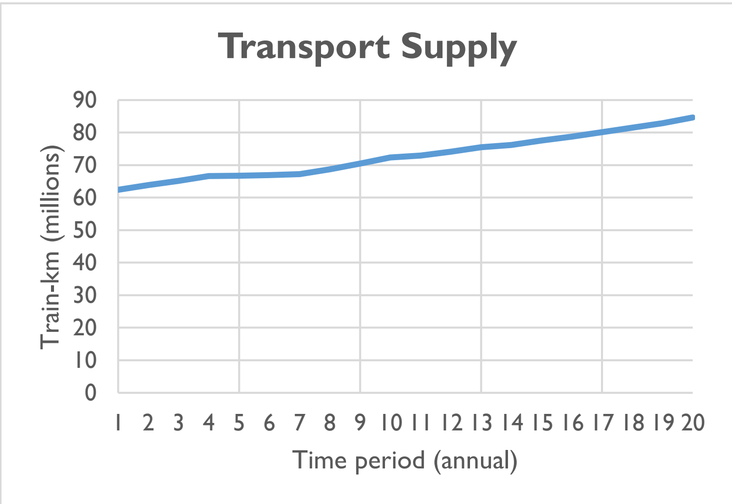

Figure 4

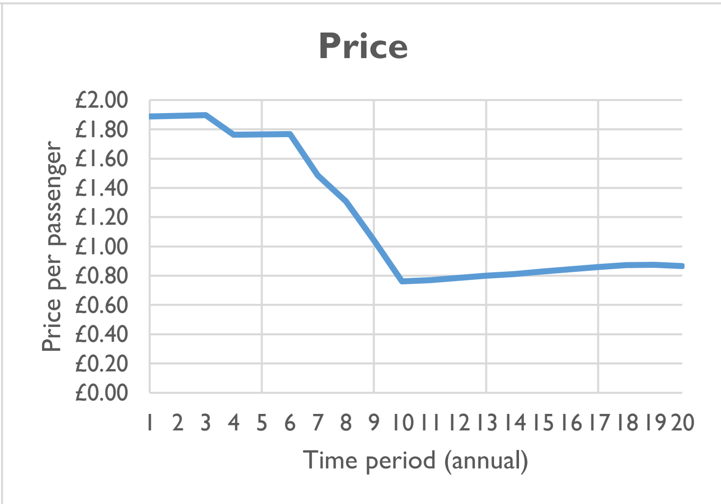

Figure 5

Conclusions (preliminary)

Given these early results, it can be seen that asset condition (and associated spending) does play a significant role in the optimisation of welfare, supply and price. In particular, the idea of an initial period of ‘correction’ before reaching an optimal steady state is also common in the infrastructure asset management literature. The gradual ‘consumption’ of A_t demonstrates the importance of the planning horizon as well as the terminal condition of the state variable, a feature of dynamic programming. Further exploring the planning horizon, and assumptions about the useful life of assets, could yield valuable insights. Finally, note that the steady-state level of A_t will depend on the assumptions and relationships built into the model, especially the interplay between renewals and maintenance, so these links between different variables will continue to be closely examined and refined to be more realistic.

References

American Society of Civil Engineers. (2017). 2017 Infrastructure Report Card: Economic Impact. Retrieved from https://www.infrastructurereportcard.org/the-impact/economic-impact/

Coen, R. (1975). Investment behavior, the measurement of depreciation, and tax policy. *American Economic Review, 65*(1), 59-74.

Globalization and World Cities (GaWC) Research Network. (2016). The World According to GaWC 2016. Retrieved from http://www.lboro.ac.uk/gawc/world2016.html

Hörcher, D., Graham, D. J., & Anderson, R. J. (2017). Crowding cost estimation with large scale smart card and vehicle location data. *Transportation Research Part B: Methodological, 95*, 105-125.

Mohring, H. (1972). Optimization and scale economies in urban bus transportation. *American Economic Review, 62*(4), 591-604.

Parry, I., & Small, K. (2009). Should urban transit subsidies be reduced? *American Economic Review, 99*(3), 700-724.

Proost, S., & van Dender, K. (2008). Optimal urban transport pricing in the presence of congestion, economies of density and costly public funds. *Transportation Research Part A*, *42*(9), 1220-1230.

Railway and Transport Strategy Centre, & World Bank. (2018, 20th March). *The Operator's Story: How Metro Operators and their Authorities can Deliver High-Quality, Sustainable Metros.* Presentation at Asia Pacific Rail, Hong Kong.

Sinha, K., Labi, S., & Agbelie, B. (2017). Transportation infrastructure asset management in the new millennium: continuing issues, and emerging challenges and opportunities. *Transportmetrica A: Transportation Science*, *13*(7), 591-606.

Small, K., & Verhoef, E. (2007). *The Economics of Urban Transportation*: Routledge.

Spy Pond Partners LLC, AECOM, McCollom Management Consulting Inc., Cohen, H., & Silkunas, S. (2018). *The Relationship Between Transit Asset Condition and Service Quality (TCRP Research Report 198)*. Transportation Research Board.

Sun, Y., Guo, Q., Schonfeld, P., & Li, Z. (2016). Implications of the cost of public funds in public transit subsidization and regulation. *Transportation Research Part A, 91*, 236-250.

Tirachini, A., Hensher, D., & Rose, J. (2013). Crowding in public transport systems: Effects on users, operation and implications for the estimation of demand. *Transportation Research Part A: Policy and Practice, 53*, 36-52.