Correlated Preferences in Spatial Equilibrium: Evidence from the Elizabeth Line

Jan David Bakker — Bocconi University
Nikhil Datta — University of Warwick
Simon Fuchs — Atlanta Fed
Amrita Kulka — University of Warwick

May 2025

1 Extended abstract

Many government interventions are spatially targeted, either explicitly—through infrastructure projects and regional transfers—or implicitly, by supporting industries that are unevenly distributed across space. Evaluating the welfare and distributional consequences of such place-based policies requires understanding how individuals substitute across locations in response to local shocks or policy changes. While canonical spatial models allow for rich interactions between locations through commuting, migration or trade, they often impose restrictive substitution patterns on worker preferences.

In this paper, we use novel data to provide the first evidence on the empirical substitution patterns across residential locations and quantify how important these are for the effects of standard infrastructure improvements through the lens of quantitative spatial models. We leverage a unique dataset on fine-grained individual searches for residential properties across locations from the UK to show that the preferences deviate strongly from the common statistical benchmarks and are heterogeneous along various dimensions across space. Based on this novel data we plan to estimate rich substitution patterns across locations and introduce them into an otherwise standard urban model.

We then use exogenous variation from the opening of the Elizabeth line in London in 2022 to eval-

uate the importance of these substitution patterns. First, we plan to test whether certain restrictions implied by the standard model that we can test using regression analysis are rejected by the data. We will then use recent advances in testing quantitative spatial and trade models – the methodology developed by Adão, Costinot, and Donaldson (2025) – to test whether allowing for richer location preferences quantitatively matters for model fit. Lastly, we aim to compare the predictions of these different models for the aggregate and distributional effects of transport infrastructure investment in order to quantitatively evaluate the importance of the different substitution patterns for aggregate outcomes.

The remainder of this paper is organized as follows. Section 2 introduces the data and Section 3 discusses descriptive patterns in the search data. In Section 4 we introduce a quantitative spatial model with a general GEV structure as well as a number of special cases: The standard Frechet-IIA model, and two models with correlated error structures. Section 5 discusses the institutional details of the Elizabeth line, our empirical setting, and Section 6 contains the current reduced-form analysis of the effects of the Elizabeth line on local rental prices and will contain the structural regression analysis derived from the models in Section 4. Section 7 contains a sketch of the Adão et al. (2025) model test and how it relates to the quantitative models outlined before.

2 Data

We combine a number of datasets for the reduced-form and structural estimation, which we document below.

First, we make use of a unique search and listings dataset of the UK housing market obtained from a very large housing market platform. The platform hosts the near universe of rental and sales residential property listings for the UK, and is one of the primary platforms used for housing search. The search data includes essentially every action taken by a user on the platform between 2019 and 2024. This includes every search, which is made up of the location term (and matching polygon) along with all other filters. Filters include minmium and maximum prices, minimum and maximum number of bedrooms, property type (e.g. detatched, flat), whether furnished (if a rental) and "must haves" (e.g. garden or parking). We additionally know which properties (identified via a unique listing-id) were viewed within that search and for how long, whether users "saved" the listing for viewing again later, and whether they initiated a contact with the associated broker. Each user is assigned a unique user-hash which we also observe, and thus we can see searches, views and contacts repeatedly over time for the same user as every single action has an associated date-time stamp. In total the sales search data contains 16.7 billion unique searches, 14.9 billion views, 100 million saves, and 17.8 million contacts, covering active user time totaling 77,000 user-years.

Similarly, the rental search data contains 3.9 billion unique searches, 3.2 billion views, 40 million saves, and 48 million contacts, covering active user time totaling 17,000 user-years.[1]

The associated listing data contains information on all sales and rental listings between 2010-2024. The data contains all information on the platform related to the property. This includes the precise address, latitude and longitude, property characteristics such as property type, floor space and number of bedrooms, the listing price and how it changed over time, the date of the initial advertisement, the "status" of the listing (i.e. available, under offer, sold), the date which the status changed, the broker and the full textual property description on the platform.

We then complement the housing search data with travel time data from the Travel Time API, between all property listings and all TfL (Transport for London) train or tube stations. This allows us to calculate how travel times between properties and TfL stations changed as a result of the Elizabeth Line. We furthermore use the public travel time matrices (TTMs) from the Urban Big Data Centre for 2021 and 2023 which calculate pairwise lower-super output area (LSOA) travel times. These are can in turn be combined with local LSOA wage data from the Annual Survey of Household Earnings for the construction of market access measures. Lastly, we use origin-destination commuting data at the LSOA level from the 2021 Census along with TfL's NUMBAT origin-destination dataset. LSOAs are the second smallest census area in the UK, and generally comprised of a population of 1,500. There are approximately 40,000 in the UK.

Combining these datasets provides allows us to identify workers' preferences for locations while accounting for the rich economic interactions in and around the Greater London area.

3 Descriptive statistics

3.1 Correlated Housing Search

In this subsection we present a number of descriptive statistics on the fine grained spatial correlations in search patterns from the housing search data. We aggregate searches to the LSOA level, the second smallest census geography in the UK (above Output Areas), which generally comprise a population of 1,500 similar to a Census block group in the US. There are approximately 40,000 LSOAs in the UK.

The correlations are based off simple conditional probabilities which give the probability of individual i searching in some LSOA g, given then have searched in another LSOA g'. When searching

[1] We carry out a number of cleaning tasks to remove bots. This includes removing those users who spend less than 1 minute on the platform, as well as those that carry out large number of searches and views in quick succession.

on the platform searchers typically do not search for a particular LSOA, but rather village, town and city names, postal areas, and landmarks such as train stations, with a selected radii. We spatially merge the polygons for these search locations to LSOAs.

Let $S_i = \{s_{il}\}$ denote the set of location-specific searches made by individual i in the set of individuals $\mathcal{I}$, where each s_{il} corresponds to a search in location l (e.g., a town, postcode, or landmark). Since these locations typically span multiple LSOAs, we define an indicator function for whether individual i searched in a given LSOA g as:

$$1[i \text{ searched } g] = \begin{cases} 1 & \text{if } \exists\, l \text{ such that } g \in l \text{ and } s_{il} \in S_i \\ 0 & \text{otherwise} \end{cases}$$

Then the conditional probability is given by:

$$\Pr(i \text{ searched } g \mid i \text{ searched } g') = \frac{\sum_{i \in \mathcal{I}} 1[i \text{ searched } g] \cdot 1[i \text{ searched } g']}{\sum_{i \in \mathcal{I}} 1[i \text{ searched } g']} \tag{1}$$

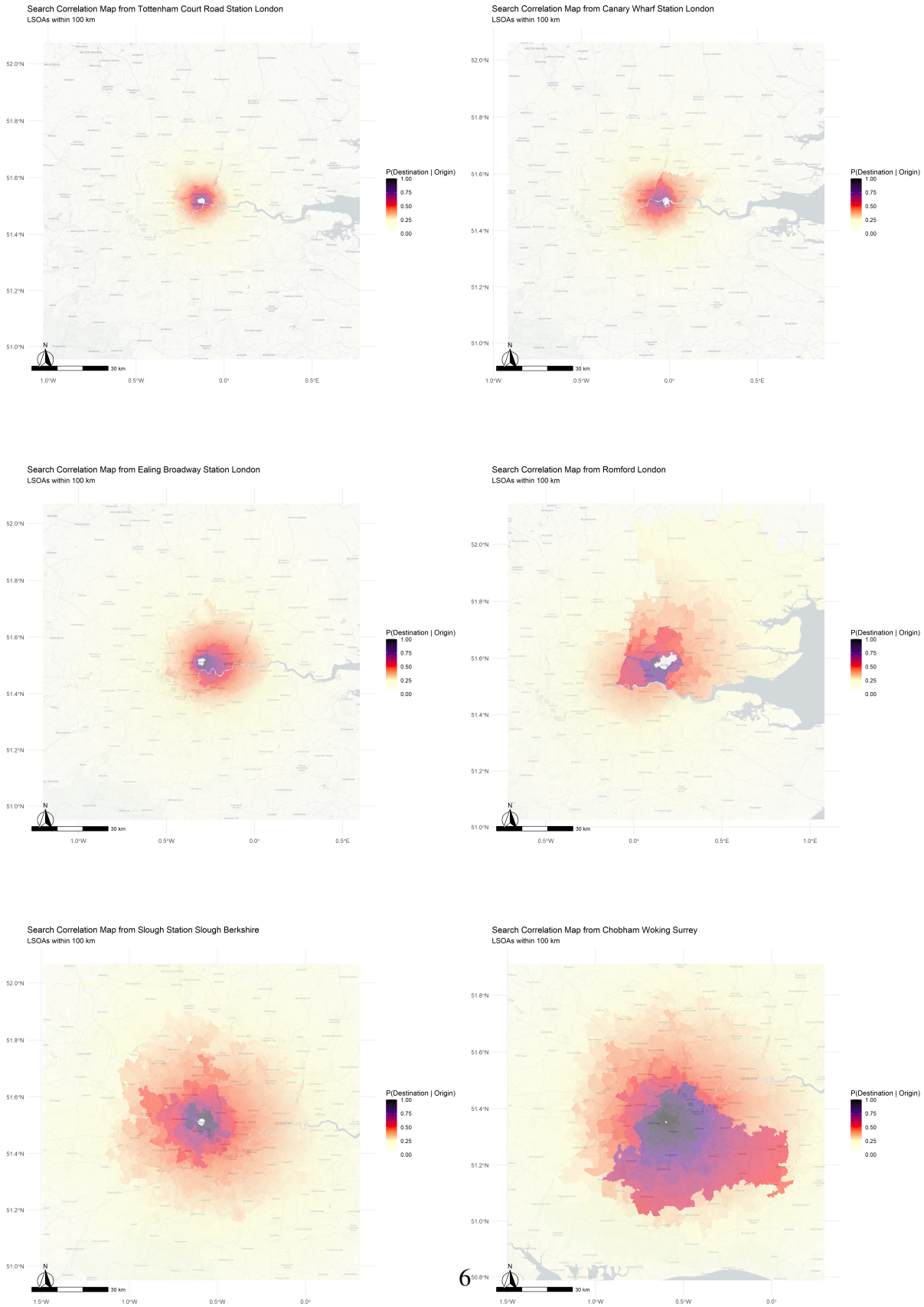
Figure 1 presents 6 spatial correlation maps, each $100km^2$, with a different LSOA, for the rental market in 2019. The white area in the centre of each map represents the search area, while the shading represents the spatial correlation in search patterns. Darker areas represent stronger correlations, while lighter represent weaker. The maps are ordered (from left to right) in ascending distance from central London. The first four maps' origin location are within the Greater London area, while the last two are areas within the Home Counties.[2]

The maps illustrate a number of interesting patterns in the data. First, the gravity coefficient is highly heterogeneous, and varies across all origin points. The spatial correlation decays much quicker, the closer one is to central London, and increases as one moves from central, to the outer parts of Greater London, and on to the Home Counties. Second, the spatial correlations are not symmetric around the origin point. For example, in the first map, which is 1 mile around Tottenham Court Road station, the spatial correlations are stronger north of the river Thames than the south. Similarly, in the second and third maps, which are centred around Canary Wharf (in the east of London) and Ealing Broadway (in the west of London) stations, the spatial correlations are stronger to the west and east of them respectively, towards central London. Third, spatial decay in correlated preferences are non-linear and non-monotonic. For example, in the third and fifth maps, centred around Ealing Broadway and Slough, respectively, both areas show stronger correlations

[2]The Home Counties are the seven counties that surround Greater London, known for being commuter areas.

for areas farther to the north than other areas closer by. For Ealing Broadway this is particularly clear along the M1 and A1M corridors (major roadways leading into London) and and for Slough towards Chesham and Amersham, the two farthest stops on the London Undergound Metropolitan Line. Lastly, they may be reflective of other aspects of idiosyncratic preferences. In particular, the last map where the origin areas are LSOAs within a small 4000 person village, Chobham, show that despite being a similar distance to London as Slough, and only a 10 minute drive to Woking train station (which has a fast commuter train into London), searchers' preferences are far more correlated with areas outside of greater London and are quite strong across much of the more rural areas to the south of London.

Figure 1: Correlated search in space



Notes: These figures provide different examples for the correlated search patterns in space.

4 Model

In this Section we develop a quantitative urban spatial model that allows for general correlational patterns in origin-route-destination choice of workers. We first develop the generic model (adapting Lind and Ramondo (2023) to an urban setting with residential location, commuting and routing choice) and then show how it collapses to the standard QSM model (Ahlfeldt, Redding, Sturm, and Wolf, 2015; Allen and Arkolakis, 2022). We then present two extensions that generate more flexible correlation structures and outline their testable implications.

4.1 Generic model

Foundations and indirect utility. A resident who lives in location $o \in \mathcal{N} = \{1, \ldots, N\}$ may take a job in location $d \in \mathcal{N}$. To reach the job she chooses a feasible route $r = (r_0 = o, r_1, \ldots, r_K = d)$ drawn from the set $\mathcal{R}_{od}$. Hence the single discrete alternative under consideration is the triple (o, d, r). Given consumption C and housing H the worker's period utility is $u_o\, C^{1-\gamma} H^{\gamma} \varepsilon_{odr}$ with $\gamma \in (0, 1)$. The budget line, $C + R_o H = w_d / \prod_{l=1}^{K} t_{r_{l-1} r_l}$, allocates the fraction γ of disposable income to housing. Substituting the budget into utility delivers the *multiplicative* indirect utility

$$V_{odr} = \Big[\frac{u_o w_d}{R_o^{\gamma} \prod_l t_{r_{l-1} r_l}} \Big] \varepsilon_{odr},$$

where the bracketed term is the systematic component $y_{odr} = u_o w_d \big(R_o^{\gamma} \prod_l t_{r_{l-1} r_l} \big)^{-1}$.

Additive-GEV representation (link to Lind and Ramondo, 2023, App. A, Lemma A-1). Taking logs converts the multiplicative form into an additive one,

$$\tilde{U}_{odr} = \ln y_{odr} + \tilde{\varepsilon}_{odr}, \qquad \tilde{\varepsilon}_{odr} = \ln \varepsilon_{odr}.$$

Because a Type–II Frechét shock becomes Gumbel after logging, the vector $\tilde{\boldsymbol{\varepsilon}}$ is an *additive* GEV error with a homogeneous-of-degree-one generating function $G : \mathbb{R}_+^J \to \mathbb{R}_+$ (homogeneity is Result A-2 in Lind–Ramondo).

McFadden choice probability for the full triple (McFadden, 1977). With $\mathbf{y} = \{y_{odr}\}$ and $G_{odr} \equiv \partial G / \partial y_{odr}$,

$$P_{odr} = \frac{y_{odr}\, G_{odr}(\mathbf{y})}{G(\mathbf{y})}.$$

Inclusive value and commuting flows. Summing P_{odr} over routes that link the same origin and destination defines

$$\Phi_{od} = \sum_{r \in \mathcal{R}_{od}} y_{odr}\, G_{odr}(\mathbf{y}), \qquad \pi_{od} = \frac{\Phi_{od}}{G(\mathbf{y})}.$$

With total employment $\bar{L}$ the commuting flow on pair (o, d) is $L_{od} = \bar{L}\,\pi_{od}$.

Market environments. Job productivity at a destination increases with crowding according to $A_d = \bar{A}_d\,(L_d^F)^\alpha$ where $L_d^F = \sum_o L_{od}$ and $\alpha \geq 0$. Amenities $u_o = \bar{u}_o$ are taken as exogenous. Housing at origin o clears when

$$\gamma \sum_d L_{od} \frac{w_d}{\Phi_{od}} = R_o\,\bar{H}_o.$$

Ex-ante welfare (McFadden log-sum formula). Additive GEV errors imply

$$\mathbb{E}\big[\max_{d,r} \tilde{U}_{odr}\big] = \gamma_E + \ln G(\mathbf{y}),$$

so after dropping the constant γ_E and exponentiating,

$$W_o = \frac{u_o}{R_o^\gamma}\, G\Big(\{z_{dr}\}_{d,r}\Big), \qquad z_{dr} = w_d \Big/ \prod_l t_{\,r_{l-1}r_l}.$$

The first factor captures real consumption at home, whereas the generating function compresses the entire wage–route landscape into a scalar that fully respects the assumed error correlation structure.

How special cases are obtained. Choosing a particular generating function closes the model algebraically. For instance the independence assumption corresponds to $G(\mathbf{y}) = \big[\sum y_{odr}^{1/\theta}\big]^\theta$, recovering the multinomial-logit formulas in Ahlfeldt et al. (2015) and Allen and Arkolakis (2022). Replacing that G by a correlated generator, such as the -square-root or the Daly–Bierlaire pairwise-combinatorial form, changes only the definition of Φ_{od} and thus propagates mechanically through flows, rents and welfare. All other equilibrium blocks remain intact because they rely only on the generic sequence "probabilities $\Longrightarrow$ inclusive value $\Longrightarrow$ flows …" established above.

4.2 Benchmark model: IIA

Independence assumption and generating function. To obtain the benchmark closed-form formulas we assume every idiosyncratic shock is *independent*. In the GEV framework this means the generating function factorises as

$$G(\mathbf{y}) = \Big[\sum_{o,d,r} y_{odr}^{1/\theta}\Big]^\theta, \qquad \theta > 0,$$

where the systematic utility component $y_{odr} = u_o w_d \big(R_o^\gamma \prod_l t_{\,r_{l-1}r_l}\big)^{-1}$ is inherited unchanged from the generic set-up.

Choice probability for a full (o, d, r) triple. With this generator the partial derivative equals

$$G_{odr} = S^{\theta-1}\, y_{odr}^{1/\theta-1}, \qquad S \equiv \sum_{k,\ell,s} y_{k\ell s}^{1/\theta},$$

so McFadden's formula gives the familiar multinomial-logit share

$$P_{odr} = \frac{y_{odr}^{1/\theta}}{\sum_{k,\ell,s} y_{k\ell s}^{1/\theta}}.$$

Inclusive value for an origin–destination pair. Summing the triple probabilities over all routes that connect the same origin and destination collapses to

$$\Phi_{od} = \sum_{r\in\mathcal{R}_{od}} y_{odr}^{1/\theta} = u_o^{1/\theta} w_d^{1/\theta} R_o^{-\gamma/\theta} \sum_{r\in\mathcal{R}_{od}} \Big(\prod_{l} t_{r_{l-1}r_l}\Big)^{-1}.$$

For later use it is convenient to factor out fundamentals and name the purely technological component

$$\tau_{od} \equiv \Big[\sum_{r\in\mathcal{R}_{od}} \prod_{l} t_{r_{l-1}r_l}^{-\theta}\Big]^{-1/\theta}, \qquad \Phi_{od} = (u_o w_d R_o^{-\gamma})^{1/\theta}\, \tau_{od}^{-\theta}.$$

Probability of an origin–destination pair and commuting flows. Because $\pi_{od} = \Phi_{od}/G(\mathbf{y})$, total commuters on the link $o \to d$ are

$$L_{od} = \bar{L}\,\frac{\Phi_{od}}{G(\mathbf{y})} = \tau_{od}^{-\theta}\, u_o^{\theta} w_d^{\theta} R_o^{-\gamma\theta} \frac{\bar{L}}{\Big[\sum_{k,\ell} \tau_{k\ell}^{-\theta} u_k^{\theta} w_\ell^{\theta} R_k^{-\gamma\theta}\Big]},$$

where the denominator is simply the value of the generator raised to the power $1/\theta$.

Labour-market clearing and agglomeration. Total residents and job-holders satisfy $L_o^R = \sum_d L_{od}$ and $L_d^F = \sum_o L_{od}$. Destination productivity follows $A_d = \bar{A}_d\,(L_d^F)^{\alpha}$ and amenities remain $u_o = \bar{u}_o$. Substituting the flow expression and renormalising with $l_o^R = L_o^R/\bar{L}$ and $l_d^F = L_d^F/\bar{L}$ produces

$$l_o^R = \chi\, \bar{u}_o^{\theta}\, R_o^{-\gamma\theta} \sum_d \tau_{od}^{-\theta}\, \bar{A}_d^{\theta}\, (l_d^F)^{\alpha\theta}, \qquad (l_d^F)^{1-\alpha\theta} = \chi\, \bar{A}_d^{\theta} \sum_o \tau_{od}^{-\theta}\, \bar{u}_o^{\theta}\, R_o^{-\gamma\theta},$$

with the scaling factor $\chi = (\bar{L}^{\alpha}/\bar{W})^{\theta}$ and $\bar{W} = \big[\sum_{k,\ell} \tau_{k\ell}^{-\theta} u_k^{\theta} w_\ell^{\theta} R_k^{-\gamma\theta}\big]^{1/\theta}$.

Housing equilibrium and expected utility. Housing clears whenever

$$\gamma \sum_{d} L_{od} \frac{w_d}{\Phi_{od}} = R_o \bar{H}_o,$$

and expected utility obeys the log-sum formula

$$W_o = \frac{u_o}{R_o^{\gamma}} \Big[\sum_{d} \tau_{od}^{-\theta} w_d^{\theta} \Big]^{1/\theta} = \bar{W}.$$

Rent equation and testable implication. Combining commuting flows with the housing condition gives

$$R_o = \Big[\frac{\gamma}{\bar{H}_o} \frac{\bar{L}}{\bar{W}^{\theta}} \bar{u}_o^{\theta} \, \mathrm{CMA}_o \Big]^{1/(1+\gamma\theta)}, \qquad \mathrm{CMA}_o = \sum_{d} \tau_{od}^{-(1+\theta)} w_d^{1+\theta}.$$

Expressing the relationship in logarithms produces the estimating equation

$$\ln R_{ot} = \gamma_o + \gamma_t + \frac{1}{1+\gamma\theta} \ln \mathrm{CMA}_{ot},$$

which is the exact rent–market-access specification implemented in Ahlfeldt et al. (2015) and Allen and Arkolakis (2022). The derivation above highlights that the result is a direct consequence of adopting the IIA generating function, making it clear how any deviation from independence will propagate through Φ_{od} and change the final empirical specification.

4.3 Correlated preferences model (1): Pairwise Correlated Logit

Correlation structure and generating function. The only departure from the independent-errors benchmark is that the idiosyncratic shocks for two routes connecting the *same* origin–destination pair are allowed to covary (Wen and Koppelman, 2001). This correlation is captured by augmenting the IIA generator with a square-root term,

$$G(\mathbf{y}) = \sum_{o,d,r} y_{odr} \; + \; \rho \sum_{(o,d,r)<(o,d,s)} \sqrt{y_{odr}\, y_{ods}}, \qquad 0 \le \rho < 1,$$

where the systematic component remains

$$y_{odr} = \frac{u_o w_d}{R_o^{\gamma} \prod_{l=1}^{K} t_{r_{l-1} r_l}}.$$

The first term reproduces the logit case; the second term adds a positive covariance proportional to the geometric mean of route utilities.

Choice probability for a full (o, d, r) triple. Differentiating G with respect to a single y_{odr} gives

$$G_{odr} = 1 + \frac{\rho}{2\sqrt{y_{odr}}} \sum_{\substack{s \in \mathcal{R}_{od} \\ s \neq r}} \sqrt{y_{ods}},$$

so McFadden's rule delivers the triple probability

$$P_{odr} = \frac{y_{odr}\left[1 + \dfrac{\rho}{2} \displaystyle\sum_{s \neq r} \sqrt{y_{ods}/y_{odr}}\right]}{G(\mathbf{y})}.$$

Inclusive value for an origin–destination pair. Adding these probabilities across all routes linking the same origin and destination yields

$$\Phi_{od} = \sum_{r \in \mathcal{R}_{od}} y_{odr} + \rho \sum_{\substack{r,s \in \mathcal{R}_{od} \\ r<s}} \sqrt{y_{odr}\, y_{ods}} \equiv \Psi_{od}.$$

The notation Ψ_{od} emphasises that the inclusive value now contains both the independent sum of route utilities and the extra "overlap bonus" induced by ρ.

Factoring out fundamentals. Writing the link cost product $\tau_{odr}^{-\theta} = \prod_l t_{r_{l-1}r_l}^{-\theta}$ and factoring out origin–destination fundamentals shows that

$$\Psi_{od} = u_o^{\theta} w_d^{\theta} R_o^{-\gamma\theta}\, \tilde{\tau}_{od}^{\mathrm{PCL}},$$

with

$$\tilde{\tau}_{od}^{\mathrm{PCL}} = \underbrace{\sum_{r} \tau_{odr}^{-\theta}}_{\text{IIA component}} + \rho \sum_{r<s} (\tau_{odr}^{-\theta} \tau_{ods}^{-\theta})^{1/2}.$$

The bracketed expression is a "correlated" generalisation of the standard CES route aggregator; when $\rho = 0$ it collapses to the logit sum $\sum_r \tau_{odr}^{-\theta}$.

Probability of an origin–destination pair and the new normalising constant. Dividing the inclusive value by the generator gives

$$\pi_{od} = \frac{\tilde{\tau}_{od}^{\mathrm{PCL}}}{\displaystyle\sum_{k,\ell} u_k^{\theta} w_\ell^{\theta} R_k^{-\gamma\theta} \tilde{\tau}_{k\ell}^{\mathrm{PCL}} \Big/ \left(u_o^{\theta} w_d^{\theta} R_o^{-\gamma\theta}\right)},$$

and the generator itself evaluates to

$$G(\mathbf{y}) = \sum_{k,\ell} u_k^\theta w_\ell^\theta R_k^{-\gamma\theta} \tilde{\tau}_{k\ell}^{\mathrm{PCL}} \equiv \bar{W}^\theta,$$

so $\bar{W}$ still denotes average real welfare, now computed under correlation.

Commuting flows. Multiplying probabilities by the labour force gives

$$L_{od} = \bar{L}\,\pi_{od} = \tilde{\tau}_{od}^{\mathrm{PCL}}\, u_o^\theta w_d^\theta R_o^{-\gamma\theta} \frac{\bar{L}}{\bar{W}^\theta}.$$

Labour–market clearing, agglomeration and housing. Total residents and job-holders are $L_o^R = \sum_d L_{od}$ and $L_d^F = \sum_o L_{od}$. Destination productivity follows $A_d = \bar{A}_d (L_d^F)^\alpha$ while amenities remain $u_o = \bar{u}_o$. Normalising by $\bar{L}$ and defining $\chi = (\bar{L}^\alpha / \bar{W})^\theta$ yields

$$L_o^R = \chi\, \bar{u}_o^\theta R_o^{-\gamma\theta} \sum_d \tilde{\tau}_{od}^{\mathrm{PCL}} \bar{A}_d^\theta (L_d^F)^{\alpha\theta},$$

$$(L_d^F)^{1-\alpha\theta} = \chi\, \bar{A}_d^\theta \sum_o \tilde{\tau}_{od}^{\mathrm{PCL}} \bar{u}_o^\theta R_o^{-\gamma\theta}.$$

Housing clears when

$$\gamma \sum_d L_{od} \frac{w_d}{\tilde{\tau}_{od}^{\mathrm{PCL}}} = R_o \bar{H}_o,$$

and expected utility equalisation remains

$$W_o = \frac{u_o}{R_o^\gamma} \Big[\sum_d \tilde{\tau}_{od}^{\mathrm{PCL}} w_d^\theta \Big]^{1/\theta} = \bar{W}.$$

Rent equation and empirical implication. Combining flows with the housing condition produces

$$R_o = \Big[\frac{\gamma}{\bar{H}_o} \frac{\bar{L}}{\bar{W}^\theta}\, \bar{u}_o^\theta\, \mathrm{CMA}_o^{\mathrm{PCL}} \Big]^{1/(1+\gamma\theta)}, \qquad \mathrm{CMA}_o^{\mathrm{PCL}} = \sum_d \big(\tilde{\tau}_{od}^{\mathrm{PCL}}\big)^{1-\theta} w_d^{1+\theta}.$$

Taking logarithms gives the estimating equation

$$\ln R_{ot} = \gamma_o + \gamma_t + \frac{1}{1+\gamma\theta}\, \ln \mathrm{CMA}_{ot}^{\mathrm{PCL}}.$$

Setting $\rho = 0$ eliminates overlap correlation, reduces $\tilde{\tau}_{od}^{\mathrm{PCL}}$ to the logit sum $\tau_{od}^{-\theta}$, turns $\mathrm{CMA}_o^{\mathrm{PCL}}$ into its IIA counterpart and recovers every formula in the multinomial-logit benchmark. Hence all the departures from independence are channelled cleanly through the modified route aggregator

$\tilde{\tau}_{od}^{\mathrm{PCL}}$.

4.4 Correlated preferences model (2)

Correlation structure and generating function. The second correlated benchmark introduces overlap in preferences using the "pairwise–combinatorial" generating function of Koppelman and Wen (2000). Enumerate every discrete alternative with a single letter $i \equiv (o, d, r)$ and let $y_i \equiv y_{odr}$ denote its systematic utility. The generator is written

$$G(\mathbf{y}) = \sum_{i<j} \left(y_i^{\varphi} + y_j^{\varphi}\right)^{1/\varphi}, \qquad 0 < \varphi \leq 1.$$

When $\varphi \to 1$ the expression inside parentheses tends to $y_i + y_j$ and summing over unordered pairs reproduces $\frac{1}{2}\sum_i \sum_{j \neq i}(y_i + y_j) = \sum_i y_i$, so the model nests the multinomial-logit benchmark as the special case of independent shocks.

Systematic utility. The deterministic part of each alternative remains

$$y_{odr} = \frac{u_o w_d}{R_o^{\gamma} \prod_{l=1}^{K} t_{r_{l-1} r_l}},$$

exactly as in the generic formulation.

Choice probability for a full triple. The partial derivative of the generator with respect to a single y_i equals

$$G_i(\mathbf{y}) = \sum_{j \neq i} \left(y_i^{\varphi} + y_j^{\varphi}\right)^{1/\varphi - 1} y_i^{\varphi - 1},$$

so McFadden's rule delivers

$$P_i = \frac{y_i \, G_i(\mathbf{y})}{G(\mathbf{y})} = \frac{y_i^{\varphi} \sum_{j \neq i} (y_i^{\varphi} + y_j^{\varphi})^{1/\varphi - 1}}{\sum_{a<b} (y_a^{\varphi} + y_b^{\varphi})^{1/\varphi}}.$$

Inclusive value for an origin–destination pair. Summing these probabilities over all routes that connect the same origin and destination gives

$$\Upsilon_{od} = \sum_{i \in \mathcal{I}_{od}} y_i^{\varphi} \sum_{j \neq i} \left(y_i^{\varphi} + y_j^{\varphi}\right)^{1/\varphi - 1}.$$

The object Υ_{od} is the pairwise-combinatorial analogue of the CES travel aggregator; it collapses to $\tau_{od}^{-\theta} = \sum_r y_{odr}$ when $\varphi = 1$. The probability of choosing destination d from origin o is simply $\pi_{od} = \Upsilon_{od}/G(\mathbf{y})$.

Commuting flows and re-scaling. Total commuters are $L_{od} = \bar{L}\,\pi_{od}$. To see the familiar economic fundamentals it is convenient to write $y_i^{\varphi} = (u_o w_d R_o^{-\gamma})^{\varphi}\,\tau_{odr}^{-\theta\varphi}$, where $\tau_{odr}^{-\theta} = \prod_l t_{r_{l-1}r_l}^{-\theta}$. Because every term inside Υ_{od} is homogeneous of degree φ in $(u_o w_d R_o^{-\gamma})$, one can factor out

$$\Upsilon_{od} = u_o^{\vartheta} w_d^{\vartheta} R_o^{-\gamma\vartheta}\,\tilde{\tau}_{od}^{\mathrm{PCL}}, \qquad \vartheta = \frac{1}{\varphi},$$

with

$$\tilde{\tau}_{od}^{\mathrm{PCL}} = \sum_{i \in \mathcal{I}_{od}} \tau_{odr}^{-\theta\varphi} \sum_{j \neq i} \big(\tau_{odr}^{-\theta\varphi} + \tau_{ods}^{-\theta\varphi}\big)^{1/\varphi - 1}.$$

Although the algebra looks forbidding, conceptually $\tilde{\tau}_{od}^{\mathrm{PCL}}$ is nothing more than the standard logit "sum of route utilities" decorated with a term that rewards pairs of high-utility routes for being available simultaneously. The overall generator turns into

$$G(\mathbf{y}) = \sum_{k,\ell} u_k^{\vartheta} w_{\ell}^{\vartheta} R_k^{-\gamma\vartheta}\,\tilde{\tau}_{k\ell}^{\mathrm{PCL}} \equiv \bar{W}^{\vartheta},$$

so the normalising constant continues to equal average welfare raised to the power ϑ.

Equilibrium conditions. With flows in hand all subsequent equations mirror those already derived for the independent case after replacing each $\tau_{od}^{-\theta}$ by $\tilde{\tau}_{od}^{\mathrm{PCL}}$ and every occurrence of θ by its reciprocal ϑ. Labour-market clearing becomes

$$L_o^R = \chi\,\bar{u}_o^{\vartheta} R_o^{-\gamma\vartheta} \sum_d \tilde{\tau}_{od}^{\mathrm{PCL}} \bar{A}_d^{\vartheta} (L_d^F)^{\alpha\vartheta}, \quad (L_d^F)^{1-\alpha\vartheta} = \chi\,\bar{A}_d^{\vartheta} \sum_o \tilde{\tau}_{od}^{\mathrm{PCL}} \bar{u}_o^{\vartheta} R_o^{-\gamma\vartheta},$$

with $\chi = (\bar{L}^{\alpha}/\bar{W})^{\vartheta}$. Housing clears when

$$\gamma \sum_d L_{od} \frac{w_d}{\tilde{\tau}_{od}^{\mathrm{PCL}}} = R_o \bar{H}_o,$$

and mobility keeps expected utility equal across origins,

$$W_o = \frac{u_o}{R_o^{\gamma}} \Big[\sum_d \tilde{\tau}_{od}^{\mathrm{PCL}} w_d^{\vartheta}\Big]^{1/\vartheta} = \bar{W}.$$

Rent equation and testable implication. Substituting the new flows and the housing condition gives

$$R_o = \Big[\frac{\gamma}{\bar{H}_o}\,\frac{\bar{L}}{\bar{W}^{\vartheta}}\,\bar{u}_o^{\vartheta}\,\mathrm{CMA}_o^{\mathrm{PCL2}}\Big]^{1/(1+\gamma\vartheta)}, \qquad \mathrm{CMA}_o^{\mathrm{PCL2}} = \sum_d \big(\tilde{\tau}_{od}^{\mathrm{PCL}}\big)^{1-\vartheta} w_d^{1+\vartheta}.$$

Taking logarithms yields the estimating equation

$$\ln R_{ot} = \gamma_o + \gamma_t + \frac{1}{1+\gamma\vartheta}\,\ln \mathrm{CMA}_{ot}^{\mathrm{PCL2}}.$$

Sending φ to one makes $\vartheta \to 1$ and $\tilde{\tau}_{od}^{\mathrm{PCL}} \to \tau_{od}^{-\theta}$, thereby reproducing every formula in the multinomial-logit benchmark. Thus all departures from the independent-error world appear solely through the modified route aggregator and the rescaling of exponents.

5 The Elizabeth line

The Elizabeth line is one of Britain's largest infrastructure projects, costing GBP19 billion, requiring 26 miles of new tunnels and stretching 73 miles from Reading to the west of London to Shenfield and Abbey Wood in the east. From a spatial perspective, it was estimated to bring 1.5 million new commuters within 45 minutes of central London, substantially altering the city's geography.

The idea of a cross-London rail linking the west and east ends of the city goes back as far as 1941, with further proposals in 1974 and a parliamentary bill introduced in 1991 by London Underground and British Rail that failed due to budget concerns. Today's Elizabeth line is the result of a revival of these early efforts in the early 2000s, culminating ultimately in a passing of the Crossrail Act in 2008. The main objectives were to reduce congestion and crowding of London's rail transport system by increasing capacity, to improve connectivity across the city and beyond as mentioned above, to future proof the transit system for expected employment and population growth, as well as to reduce car traffic and pollution.

Figure 2 shows how the Elizabeth line extended the existing tube network operated by Transport for London (TfL) towards the east and west, where empty circles indicate new stops to the system. Rather than building entirely new stops, the Elizabeth line combined several existing lines that were operated by other rail operators and included these in the TfL network. Many of these stops were previously part of other rail operators. Specifically, the eastern corridor from Liverpool Street Station to Shenfield Station was operated by Greater Anglia and the western route from Paddington Station to Reading Station was run by Great Western Railway before being integrated into the TfL network through the Elizabeth line. The Elizabeth line made several direct connections possible

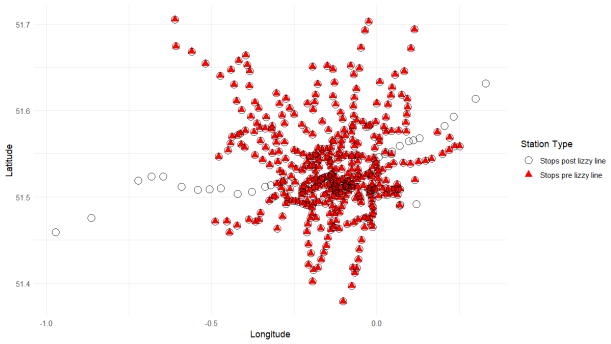


Figure 2: Tube network before and after the Elizabeth line

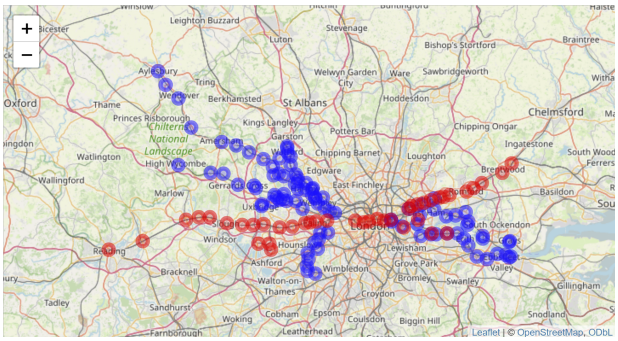


Figure 3: Alternative considered routes

that did not previously exist such as connecting the eastern and western suburbs to the center of the city and to each other and providing an easy connection between Heathrow Airport and central London. In general, while the Elizabeth line vastly improved connectivity, it did not necessarily make trips between various locations substantially cheaper, particularly so for the suburban stops.

While the route through the Elizabeth line takes through the core of the city was agreed on straightforwardly, several alternatives were considered for the eastern and western corridors. Figure **??** shows the alternative routes that were considered but ultimately not chosen after being assessed along environmental, economics, safety, accessibility, and transport integration dimensions. We will make use of these runner up locations as part of a baseline reduced-form strategy to evaluate the impacts of the Elizabeth line to be compared against model-derived alternatives.

The Elizabeth line construction began in May 2012 and different segments opened in different stages as sections transferred over to TfL. The eastern corridor from Liverpool Street to Shenfield began to run as part of TfL from June 2017, the section between Paddington and Heathrow started operating in May 2018, and the western corridor from Paddington to Reading opened in December 2019. The central section and thereby the full line opened in May 2022, delayed by four years.

6 Regression analysis

We provide preliminary motivating evidence of the Elizabeth line's impact on residential rents in two ways. First, we estimate rent gradients based on distance to the closest tube station that is part of the TfL network in the spirit of Khanna, Nyshadham, Ramos-Menchelli, Tamayo, and Tiew (2023). We estimate regressions of the following form:

$$\text{Log(Rent)}_{int} = \alpha(\text{Log(Distance to closest TfL station)}_{it} \times \text{Post}_t) + \beta' X_{it} + \theta_n + \gamma_t + \epsilon_{int} \quad (2)$$

where i is a residential unit, n is a location or neighborhood (local authority or post district) and t is a year-month. X_{it} contains unit characteristics, namely floor area and the number of bathrooms. We include varying levels of fixed effects at the neighborhood or unit level and year-month fixed effects on their own or interacted with locations. We begin by focusing only on data from 2020 to 2024, that is after the last partial opening of the Elizabeth line (western corridor) to estimate the impact of the full opening in May 2022. Future versions will incorporate the partial openings as well. Unfortunately, our data go back only to 2010, therefore we do not observe a pre period for the start of construction in 2009 and will not be able to capture overall anticipation effects. We initially consider a sample of units that lie within at most 10km driving distance of an Elizabeth line station. The sample includes units for which the distance to the closest station changes after the opening of the Elizabeth line.

Table 1 shows the estimates of α for various specifications of fixed effects and controls. As expected, being further away from a tube station is associated with lower rents post opening of the Elizabeth line. Columns 4 and 5 show specifications that control for unit-specific fixed effects, leveraging within-unit variation. The estimates imply a seemingly small elasticity. However, for units in the suburban local authorities Reading and Brentwood (where Shenfield lies) the average change in distance to TfL station post-Elizabeth line is a decline of 6.9km for Brentwood and a decline of 32.1km for Reading. These decreases in distances translate into substantial price increases of GPB189 in Reading and GBP207 in Brentwood respectively relative to average rental prices pre-opening of GBP1101 and GBP1307 respectively.

Second, we define treatment as a binary indicator, and estimate a standard distance-based differences-in-differences design (Diamond and McQuade, 2019; Blanco and Neri, 2023). We consider units as treated if they lie within 0.5km of an Elizabeth line stop following one of the empirical strategies in Gupta, Van Nieuwerburgh, and Kontokosta (2022). Control units are given by units that lie between 0.5km to 1km of an Elizabeth line stop. We then estimate the following empirical specification:

$$\text{Log(Rent)}_{int} = \alpha\text{Close to Station}_i \times \text{Post Opening}_t + \delta_1\text{Close to Station}_i + \beta' X_{it} + \theta_n + \gamma_t + \epsilon_{int} \quad (3)$$

	(1)	(2)	(3)	(4)	(5)
	Log(Rent)	Log(Rent)	Log(Rent)	Log(Rent)	Log(Rent)
Log(Distance to closest TfL station)$_{it}$ × Post$_t$	-0.0983***	-0.00830***	-0.00493***	-0.00780***	-0.00204***
	(0.000945)	(0.000716)	(0.000692)	(0.000347)	(0.000349)
Unit characteristics	Yes	Yes	Yes		
Year-month f.e.	Yes	Yes	Yes	Yes	
LA f.e.		Yes			
Post district (PD) f.e.			Yes		
Unit f.e.				Yes	Yes
LA X year-month f.e.					Yes
N	481737	481737	481736	302338	302257
R^2	0.357	0.597	0.636	0.974	0.975

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1: Rent gradient

Table 2 shows these results. Rents increase robustly across specifications by about 1.2% within half a kilometer of an Elizabeth line station relative to units within half a kilometer and a kilometer from the station amounting to GBP23. Given the dramatic decline in distance to a tube station in Reading ($\approx$ 94%), it is quite likely that a substantially larger area is treated and that even units more than a kilometer away experience price increases as a result of the new tube link, such that our estimates should be treated as a lower bound. Our next steps will involve leveraging the model-implied control areas as well as using our unique housing search data to define areas that might serve as good controls based on search preferences but without suffering from direct spatial spillovers.

	(1)	(2)	(3)	(4)	(5)
	Log(Rent)	Log(Rent)	Log(Rent)	Log(Rent)	Log(Rent)
Close to Station X Post Opening	0.0121***	0.0107***	0.0100***	0.0220***	0.0121***
	(0.00248)	(0.00205)	(0.00194)	(0.00126)	(0.00137)
Unit characteristics	Yes	Yes	Yes		
Year-month f.e.	Yes	Yes	Yes	Yes	
LA f.e.		Yes			
Post district (PD) f.e.			Yes		
Unit f.e.				Yes	Yes
LA X year-month f.e.					Yes
N	374162	374162	374161	228139	228111
R^2	0.386	0.587	0.630	0.972	0.973

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2: Distance-based specification

7 Model validation and quantitative model comparison

This section formalises how we evaluate the empirical validity of the different quantitative spatial models introduced in Section 4. Our approach follows the testing procedure proposed by Adão et al. (2025), adapted to an urban setting with infrastructure shocks.[3] The key idea is to confront *model–implied* changes in a set of welfare-relevant outcomes with the *observed* changes triggered by the Elizabeth line, while using the exogenous time–distance shock as an instrument that purges the comparison from the influence of all other shocks.

7.1 Outcomes and notation

Let g index granular spatial units (*e.g.* LSOAs) and let $y_{g,t}$ denote an outcome whose change enters the first–order approximation to aggregate welfare in the urban model:[4]

(a) $\ln R_{g,t}$: advertised monthly rent for unit g;

(b) $\ln C_{g,t}$: average cost (generalised travel time) of the cheapest route to the CBD;

(c) $\ln S_{g,t}$: share of *searches* in location g (intensive housing demand measure).

Collect them in the vector $\Delta \mathbf{y}_g \equiv (\Delta \ln R_g,\ \Delta \ln C_g,\ \Delta \ln S_g)'$, where Δ denotes the difference between the pre-opening (2019–21) and post-opening (2022–24) averages.

[3] See also Kehoe, Pujolàs, and Rossbach (2017) for an earlier survey of goodness-of-fit exercises in trade.

[4] In the trade context Adão et al. (2025) work with import prices, export prices, and import quantities; here we use the natural urban analogues.

For each model $m \in \{\text{IIA}, \text{PCL}(\rho), \text{PC}(\varphi)\}$ described in Section 4, let $\widehat{\Delta \mathbf{y}}_g^m$ be the *predicted* counterfactual change implied by feeding the observed Elizabeth-line changes in origin–destination travel times into the calibrated model (from the 2019 baseline).

7.2 First-stage instrument

Define the model-specific *inclusive-value shock*

$$z_g^m \;=\; \sum_d \Big(\frac{\partial \Phi_{gd}^m}{\partial \tau_{gd}}\Big)\big(\tau_{gd}^{\text{post}} - \tau_{gd}^{\text{pre}}\big) \;-\; \bar{z}^m, \qquad \bar{z}^m \;=\; \frac{1}{G}\sum_g z_g^m,$$

where τ_{gd} is the composite commuting cost defined in Section 4 and Φ_{gd}^m the model-specific route-inclusive value. Intuitively, z_g^m is the *predicted rent-relevant shock* for location g that is strictly exogenous because it is pinned down by engineering travel-time changes and baseline route choice shares.

7.3 IV test à la Adão et al. (2025)

For each outcome $k \in \{R, C, S\}$ estimate

$$\Delta y_{gk} \;=\; \beta_k^m \, \widehat{\Delta y}_{gk}^m \;+\; u_{gk}, \qquad \text{IV: } \widehat{\Delta y}_{gk}^m \text{ instrumented by } z_g^m, \tag{4}$$

allowing for an unrestricted constant and geography fixed effects. Under standard regularity conditions the null $H_0 : \beta_k^m = 1$ holds *if and only if* model m correctly captures the structural response of outcome k to the Elizabeth-line shock. We report robust confidence intervals using the conservative AKM variance estimator recommended by Adão et al. (2025) to accommodate many, possibly correlated, spatial units.

Joint test. Stacking all three outcomes we run the SUR analogue of (4) and test $H_0 : \beta_R^m = \beta_C^m = \beta_S^m = 1$ via a Wald χ^2 statistic.

7.4 Power comparison across models

(i) **Benchmark (IIA).** The multinomial-logit structure yields distance-based CMA shocks identical to the standard Ahlfeldt et al. (2015) regression. A rejection here would confirm that spatial substitution patterns are *not* IIA-consistent.

(ii) **Pairwise-correlated logit (PCL).** We recalibrate ρ so that the model matches the empirical cross-elasticities of search flows documented in Section 3. Power gains arise if the richer correlation structure explains *additional* variation beyond the CMA term.

(iii) **Pairwise-combinatorial (PC).** We choose φ to fit the tail of the empirical distribution of conditional search probabilities. Because substitution elasticities vary with the number of

close substitutes, this model is expected to outperform PCL in locations with many near neighbours.

7.5 Counterfactual welfare decomposition

When a model m is *not rejected* we exploit the log-sum formula (cf. Section 4.1) to decompose aggregate welfare gains from the Elizabeth line into:

$$\underbrace{\Delta \ln G^m}_{\text{direct commute benefit}} + \underbrace{\sum_o \lambda_o^m \, \Delta \ln u_o}_{\text{amenity re-sorting}} - \underbrace{\gamma \sum_o \lambda_o^m \, \Delta \ln R_o}_{\text{housing price response}},$$

where λ_o^m are model-implied residential shares and G^m the GEV generating function. The decomposition isolates how much of the total welfare gain would be missed had one imposed the restrictive IIA assumption.

8 Conclusion

[TBD]

References

Adão, R., A. Costinot, and D. Donaldson (2025). Putting quantitative models to the test: An application to the us-china trade war. *The Quarterly Journal of Economics 140*(2), 1471–1524.

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.

Allen, T. and C. Arkolakis (2022). The welfare effects of transportation infrastructure improvements. *The Review of Economic Studies 89*(6), 2911–2957.

Blanco, H. and L. Neri (2023). Knocking it down and mixing it up: The impact of public housing regenerations.

Diamond, R. and T. McQuade (2019). Who wants affordable housing in their backyard? an equilibrium analysis of low-income property development. *Journal of Political Economy 127*(3), 1063–1117.

Gupta, A., S. Van Nieuwerburgh, and C. Kontokosta (2022). Take the q train: Value capture of public infrastructure projects. *Journal of Urban Economics 129*, 103422.

Kehoe, T. J., P. S. Pujolàs, and J. Rossbach (2017, August). Quantitative trade models: Developments and challenges. *Annu. Rev. Econom. 9*(1), 295–325.

Khanna, G., A. Nyshadham, D. Ramos-Menchelli, J. A. Tamayo, and A. Tiew (2023). *Spatial mobility, economic opportunity, and crime*. Harvard Business School.

Koppelman, F. S. and C.-H. Wen (2000, February). The paired combinatorial logit model: properties, estimation and application. *Trans. Res. Part B: Methodol. 34*(2), 75–89.

Lind, N. and N. Ramondo (2023, February). Trade with correlation. *Am. Econ. Rev. 113*(2), 317–353.

McFadden, D. (1977). Modelling the choice of residential location. Technical Report 477, Cowles Foundation for Research in Economics, Yale University.

Wen, C.-H. and F. S. Koppelman (2001, August). The generalized nested logit model. *Trans. Res. Part B: Methodol. 35*(7), 627–641.