Analysing the Impact of Sports Facility Distribution on Health Outcomes in England: A Vector Data Perspective

Park, Jae Hun

University of Cambridge

May 2025

Abstract

This study examines how the structure, capacity, and spatial configuration of sports infrastructure affect obesity outcomes in England. It introduces a novel framework combining heat diffusion, sparse Principal Component Analysis (sPCA), and Latent Dirichlet Allocation (LDA) to uncover latent exposure patterns. Using 2013–2023 panel data and instrumented fixed effects regressions, it finds that gyms significantly reduce obesity—approximately a 9% drop with a one-standard deviation increase in supply. In contrast, leisure venues and outdoor facilities show negligible or adverse effects. These findings highlight the need for targeted, data-driven infrastructure policy to address persistent spatial and socio-economic health disparities.

Acknowledgment This is my undergraduate dissertation project for BA Economics in University of Cambridge.

Plagiarism Declaration I confirm that this is entirely my own work and has not previously been submitted for assessment, and I have read and understood the University's and Faculty's definition of Plagiarism.

Anonymisation of Work Declaration I confirm that I have taken all reasonable steps to ensure that all submitted files for assessment have been anonymised and do not contain any identifiable information to me.

Word Count: 7,499 - 6,349 words including equations, plus 1,150 for Tables and Figure: 100(p.4), 200(pg. 12), 100(pg.16), and 750(p.20-22) words, respectively.

1 Introduction

Motivation. The NHS faces intensifying strain from lifestyle-related conditions. As health economists have long noted (), late-stage interventions yield limited returns, while preventative strategies like physical activity offer cost-effective alternatives (Cutler & McClellan, 2001; Marmot et al., 2010; Chandra & Skinner, 2012). Yet despite £1.5 billion in investment, activity rates have risen by just 1.5% (Public Accounts Committee, 2023), exposing a provision–demand mismatch.

Literature Gap. A substantial body of work explores links between facility access and participation (Kaczynski & Henderson, 2007; An & Sturm, 2015), but results diverge. Some report strong effects (Downward & Rasciute, 2011); others find none (Nichols et al., 2015). These inconsistencies reflect structural flaws. Co-location inflates multicollinearity: sports halls make up only 11.5% of facilities but appear in 33.9% of sites, and over half of gyms share sites with studios (Sport England, n.d.-a). Many studies use arbitrary access thresholds or omit facility type and capacity (Hallmann et al., 2012; Kokolakakis et al., 2014; Downward et al., 2024), masking spatial overlap and latent exposure. GIS nesting adds further aggregation bias. Moreover, most models assume full utilisation, ignoring network structure. A new approach is needed - one that models bundled infrastructure, captures spatial diffusion, and accounts for heterogeneity.

Contribution. This study develops such a framework by integrating: (1) full facility typology, and (2) capacity depth via latent structure detection using sPCA and LDA, along with (3) distance-based spatial connectivity. The resulting components form interpretable infrastructure archetypes, **generalisable** to Active Places Dataset (APD) studies.

Empirical Strategy & Findings. Leveraging 2013–2023 panel data and IV-FE estimation, the analysis found gyms reduce obesity; Leisure facilities show adverse or weak effects. Spatial integration was associated with regional provision yet had no direct impact, underscoring the need for facility-type-specific, demand-responsive planning.

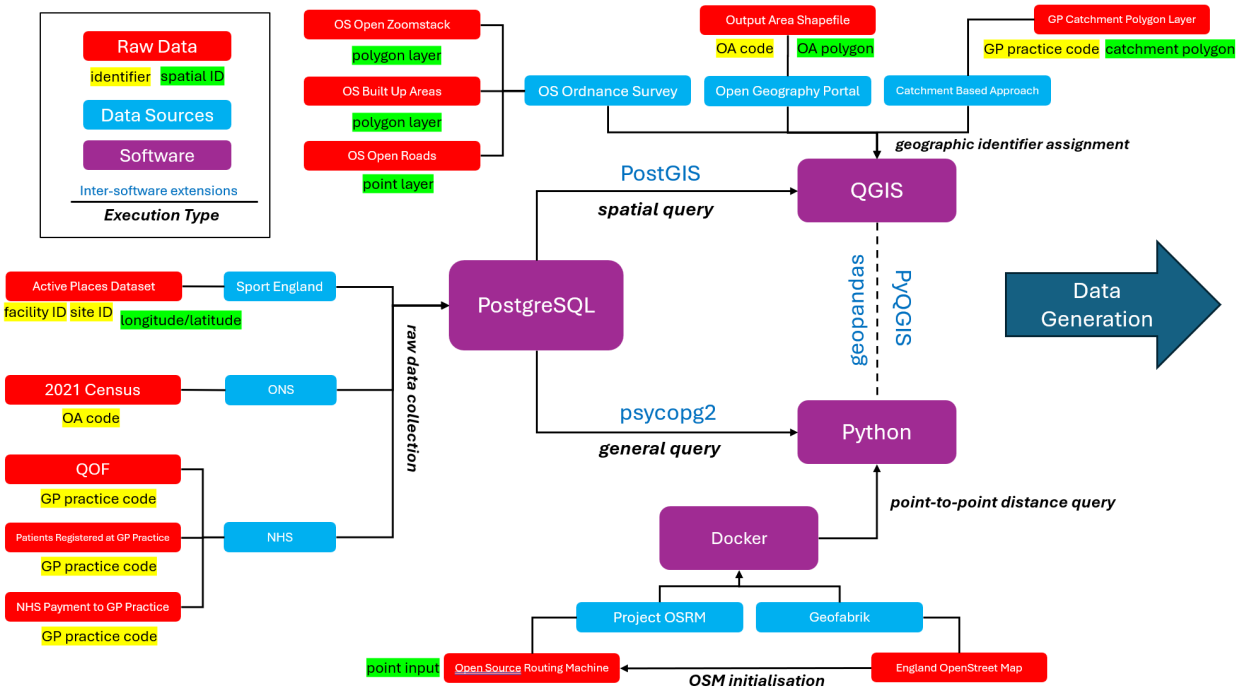

Figure 1: Data Generation Process (DGP)

2 Data

2.1 Data Generation Process

DGP (see Figure 1) projects raw inputs to the GP catchment level—the analytical unit of interest. Where lower- and upper-level identifiers are directly matched, nesting was performed via SQL queries in PostgreSQL. Otherwise, spatial joins were conducted in QGIS using polygon overlays, allowing spatially unconstrained data to be integrated into the framework. QGIS also computed polygon-derived metrics. After assigning unique identifiers, variables were nested. Data processing combined Python, PostgreSQL, and QGIS; distance computation used Open Source Routing Machine (Project OSRM contributors, n.d.) hosted on Docker. The analytical environment therefore combined structured querying, geospatial projection, and transport network modelling across PostgreSQL, QGIS, and Python.

Pre-treatment. To mitigate collinearity and reveal intra-facility variation, selected capacity measures were normalised prior to aggregation. All normalisation occurred within facility-type, avoiding artificial cross-variable correlation.

During nesting, unmatched facility-capacity combinations yield N/A values. Under the model assumption that sites select from a universal menu of facilities, these N/As represent unchosen (unrealised) capacities and are analytically equivalent to zeros. Replacing them with zero preserves the sparsity structure necessary for sparse PCA in later dimensionality reduction.

2.2 Data Sources

GP Practices GP-level obesity prevalence were obtained from Quality and Outcomes Framework (QOF; NHS Digital, 2013-2023a), which covers 97.3% of practices in England (NHS Digital, 2024a). These conditions were selected for their public health relevance and the preventative role of sports infrastructure. Unlike other lifelong or incurable conditions in the QOF (e.g., atrial fibrillation, diabetes), obesity is clinically curable and represent key physical and mental risk factors. This makes them central to evaluating supply-side public health impacts. Practice-level demographics, patient survey data and payment from NHS, were sourced from NHS registers (NHS Digital, 2013-2023b; NHS England, 2013-2023).

Sports Infrastructure APD provides detailed records of 42,461 sites and 121,200 facilities across England, structured hierarchically:

- **Site**: provision venue
- **Facility**: sport-specific installation
- **Capacity**: installation quantity

To avoid colloquial ambiguity, these levels are **strictly distinguished** in the analysis.

Urban–Rural Classification Urbanity scores were calculated using the 2021 Census classification of Output Areas (OAs; ONS, 2023), scored 0–11 by urban density. GP catchments were assigned area-weighted averages of intersecting OA scores using QGIS.

Open Source Routing Machine (OSRM) Shortest path distances were computed via OSRM on a local Docker[1] engine, using the OpenStreetMap (OSM; OpenStreetMap contributors, n.d.) of England.

GIS Layers GP catchment layer is publicly available (Catchment Based Approach, n.d.), based on patient registration patterns. Residential nodes were extracted from intersection between Built-Up Areas (BUAs; Ordnance Survey, 2022) and nodes from OS Open Roads (Ordnance Survey, 2023a). Environmental variables were derived from OS Zoomstack (Ordnance Survey, 2023b) using QGIS.

2.3 Potential Data Issues

Facility and Site Data Filtering A 10-year panel (2013/14–2022/23) was constructed from APD records. Facilities such as Athletic Halls, Cycling, and Gymnastics were excluded due to data inconsistencies or lack of coverage. Facilities were included only in years when open or operational, based on

[1]Linux backend for Windows

creation or build dates. Grass pitches (14,907) were retained continuously, reflecting ongoing use despite missing open dates[2].

GP Practices The available GP catchment layer was cross-sectional hence assumed stable over the study period. Practices with invalid geometries, incomplete polygons, or atypical units (e.g., military or institutional) were removed. Only continuously operating practices were retained, reflecting national GP consolidation (8,044 to 6,419, 2013–2023; UCL News, 2024).

Result: The final dataset retained 113,269 facilities and 39,723 sites (93.5% and 93.6% of the originals), and 6,043 practices (94.1%). Patient counts were normalised per catchment to mitigate attrition bias.

[2]This complies with APD's definition of grass pitches: "Area of grass that is marked out for at least part of the year as a pitch for a particular sport, upon which a match could be played."

3 Methodology

This section develops models that build a spatial connectivity layer to capture site-level variation. It employs distance-based heat diffusion to measure each site's effective reach and an sPCA–LDA framework to reveal latent distributional patterns in England's infrastructure market.

3.1 Heat Diffusion Model

Intuition of the Heat Diffusion Model Site utilisation declines with distance from residences and is conceptualised as 'heat' dissipating over space. Each site's influence decays with separation, modelled using exponential decay, analogous to the measure of centrality in network theory (e.g. Katz, 1953).

Residential Node Construction. Residential demand was proxied using road end nodes within each catchment's built-up area (BUA), defined as "irreversibly urban" and covering 94.9% of England's population (ONS, 2021). Degree-one nodes from OS data represent residential access and correlate strongly with BUA area ($r = 0.9837$), preserving both spatial scale and population granularity.

Distance Calculation. OSRM computed the multimodal (car, bike and foot) shortest distance between site and node, incorporating real-world travel constraints such as traffic lights from OSM, to reflect actual travel behaviour.

Model Formulation. Let d_{si} denote the road network distance between facility site $s \in S$ and residential node $i \in N$, and let $d_{1/2}$ be the half-life distance **at which a site's attractiveness decays to 50%**. The unidimensional (single-mode) heat score for site s is:

$$\text{Heat}_s = \sum_{i \in N} \exp\left(-\frac{\ln(2)}{d_{1/2}} \cdot d_{si}\right)$$

To incorporate multimodal access, let $M = \{\text{car}, \text{bike}, \text{foot}\}$ and define:

$$\alpha_m = \frac{\ln(2)}{d_{1/2}^{(m)}} \quad \text{where} \quad d_{1/2}^{(m)} \in \{5000 \text{ m (car)},\ 2500 \text{ m (bike)},\ 1200 \text{ m (foot)}\}$$

Then, the mode-averaged spatial integration score becomes:

$$\text{Heat}_s = \frac{1}{|M|} \sum_{m \in M} \sum_{i \in N} \exp(-\alpha_m \cdot d_{si}^{(m)})$$

Integration Variables. For a catchment c with site set $S_c \subseteq S$ and node set $N_c \subseteq N$, Define the heat-based connectivity metrics as:

$$\text{Integration}_c^{\text{site}} = \frac{1}{|N_c||S_c|} \sum_{s \in S_c} \text{Heat}_s$$

which captures the overall accessibility of sites to residents within the catchment.

Robustness and Decay Choice. The model was robust to $\pm 20\%$ adjustments in $d_{1/2}^{(m)}$, with no significant change in estimates. While alternative decay forms (e.g., logarithmic or step-wise) may better reflect sharp behavioural cut-offs, they require complex, area-specific calibration with limited empirical support. The exponential decay used here offers a smooth, generalisable fit, in line with Sport England's Facility Planning Model (FPM; Sport England, n.d.-b) and UK travel behaviour evidence (Langford et al., 2019; Higgs & Gilleard, 2015).

3.2 sPCA–LDA Framework

Intuition of the sPCA–LDA Framework To bridge the observational gap in the latent distributional pattern, facility provision was framed as a constrained optimisation problem to capture the mechanism. Estimation of the equilibrium instance of the formal[3] model takes into account the previously disregarded heterogeneity. The model revealed mathematical properties that allows sPCA to uncover dominant patterns in site composition, robust to noise. Their alignment with topic clusters from LDA validates their semantic coherence—akin to how stylised facts guide empirical macro models using compact summaries of real-world structure. The below are the foundational informed facts about England's sport infrastructure market and modelling assumptions[4].

Fact 1 (Imperfect Market Structure)**.** *The sports infrastructure market is imperfect, with some providers holding significant market power, characterised by economies of scope and high fixed costs (Coalter, 2013; Kokolakakis et al., 2012; Sport England, 2022).*

Fact 2 (Partitioned Preference)**.** *Consumer preference is driven by discrete, segmentable factors such as gender, social motivation, and activity preferences (Green, 2008; Eime et al., 2013; Allender et al., 2006; Tsou et al., 2022).*

Assumption 1 (Rational Expectation)**.** *Firms and consumers optimise current decisions based on anticipated future prices, demand, and outcomes, using all available information.*

Assumption 2 (Full Opportunity Set)**.** *Each site has access to all facility types, with realised site-capacity resulting from site-level constrained optimisation.*

Assumption 3 (Finite Business Types)**.** *The market has a finite number of business types (limited combinations of individual capacities).*

Assumption 4 (Finite Demand Profiles)**.** *The market has a finite number of demand profiles (limited combinations of preferences).*

Assumption 5 (Brand Signal)**.** *Businesses reveal their facility profiles through their names.*

Assumption 6 (Bag-Of-Words)**.** *Business identity is determined by word inclusion, not word order.*

[3]Sketch proof in Appendix 7.2.
[4]Justification in Appendix 7.1.

3.2.1 Stylised Model

Define the cost and utility structure as:

Consumer. Consumers choose $\mathbf{x} \in \mathbb{R}^n_+$ to maximise:

$$\max_{\mathbf{x}\geq 0} \left\{ \mathbb{E}_t \left[\sum_{i=1}^{n} u_i(x_i) \right] - x_i \mathbb{E}_t[p_i] \right\} \quad \text{s.t.} \quad \sum_{i=1}^{n} x_i \mathbb{E}_t[p_i] \leq m$$

$$u_i(x_i) = \begin{cases} U_i(x_i), & \text{if } i \in P \subseteq \{1, \ldots, n\} \\ 0, & \text{otherwise} \end{cases}$$

Firm. Firms choose $\mathbf{x} \in \mathbb{R}^n_+$ to maximise:

$$\max_{\mathbf{x}\geq 0} \mathbb{E}_t \left[\sum_{i=1}^{n} x_i p_i - \sum_{i=1}^{n} c(x_i) - \sum_{G\in\mathcal{G}} S_G \cdot \mathbb{1} \left(\sum_{j\in G} \mathbb{1}(x_j > 0) > 0 \right) \right]$$

$$c(x_i) \text{ is increasing and convex,}$$
$$\mathcal{G} = \{G_1, \ldots, G_K\} \text{ partitions } \{1, \ldots, n\},$$
$$S_G \text{ is a fixed cost for activating group } G.$$

S_G formalises Fact 1 because *additional* installation of capacities within the same scope can be done without duplicate fixed cost - i.e. economies of scope. $u_i(x_i)$ formalises Fact 2 where consumers are indifferent unless the nature of facility associated with capacity x suits their taste. The optimisation becomes:

Lemma 1 (Partitioned Sparsity and Archetype Representation)**.**

Let:

1. $\mathcal{G} = \{G_1, \ldots, G_K\}$ *be a set of facility groups,*
2. *let* $\mathbf{x}_i^*$ *be the profit-maximising capacity vector for region* i*.*

Then:

1. $\mathbf{x}_i^* = \bigoplus_{k\in K_i^*} \mathbf{x}_{G_k}^*$, *with* $x_j = 0$ *for all* $j \notin \bigcup_{k\in K_i^*} G_k$, *where* $K_i^* \subseteq \{1, \ldots, K\}$ *indexes active groups.*
2. *Under Assumptions 3 and 4, this implies:* $\mathbf{x}_i^* = \sum_{j=1}^{K} \alpha_j^i m_j$, $\alpha_j^i \geq 0$ *for a finite set* $M = \{m_1, \ldots, m_K\}$ *of partitioned sparse archetypes.*
3. $\mathbf{x}_i^*$ *fully characterised by the selection and scaling of archetypes* $m_j \in M$*.*

Interpretation. Each equilibrium catchment-level capacity vector is a modular aggregation of sparse latent business archetypes with only a subset of facility groups active.

3.2.2 Estimating Archetypes via sPCA

Context. sPCA naturally clusters nonzero capacities while enforcing sparsity, matching the latent archetype in Lemma 1. Hence the archetypes are estimated by the below from Witten et al. (2009):

Algorithm 1 (sPCA Archetype Estimation)**.**

1. ***Input**: Scaled, **uncentred**[5] site-capacity matrix $\mathbf{X} = [\mathbf{x}_1^*, \ldots, \mathbf{x}_N^*] \in \mathbb{R}^{n \times N}$, and the **sparsity constraint** $\rho \in [1, \sqrt{\dim(col(\mathbf{X}))}]$*

2. ***Solve**: $\mathbf{v}_{(i)} = \arg\max_{\mathbf{v}} \left\{ \mathbf{v}^\top \mathbf{X}^\top \mathbf{X} \mathbf{v} \right\}$ s.t. $\|\mathbf{v}\|_2 \leq 1,\ \|\mathbf{v}\|_1 \leq \rho$*

3. ***Output**: Sparse loading vectors $\mathbf{v}_{(1)}, \ldots, \mathbf{v}_{(K)}$, serving as **candidate** partitioned sparse archetypes $M = [m_1, \ldots, m_K]$.*

***Tuning of** ρ: 10-fold cross-validation with "min$\{SPE\}$ within one-standard-deviation rule" (Guerra-Urzola et al., 2021), implemented in the PMA package (Witten et al., 2024) in R.*

***Sign Restrictions**: PMA enables restricting $\mathbf{v}_i \subset \mathbb{R}^+$ or $\mathbb{R}^-$, avoiding ambiguity where opposite signs would represent both different archetypes and scale effect.*

Lemma 1 requires strictly partitioned-sparsity, which is empirically unrealistic. Thus, estimation applying Algorithm 1 on APD requires *approximate* sparsity by the following:

Lemma 2 (Approximate sPCA structure)**.**

*Let: $sPC_1, \ldots, sPC_K$ be the **Output** from Algorithm 1*

If:

1. *the first K' components have* approximately disjoint *supports, such that*

$$\forall\, a \neq b \leq K' \quad \frac{|supp(sPC_a) \cap supp(sPC_b)|}{\min(|supp(sPC_a)|, |supp(sPC_b)|)} \leq \delta \quad (\delta > 0 \text{ small}),$$

2. *Components $k > K'$ exhibit greater overlap and lack clear partitioning -i.e. the condition does not hold for most $a \neq b \geq K'$.*

[5] Appendix 7.3 details justification.

Then: $\{sPC_1, \ldots, sPC_{K'}\}$ *approximate modular archetypes* $m_k^*(\theta_i)$*, while* $\{sPC_{K'+1}, \ldots, sPC_K\}$ *reflect local adjustments, or residual noise.*

Lemma 2 **relaxes** Lemma 1 to allow small overlaps. The δ is determined empirically from leading components, providing an upper bound of disjointness for the chosen K'. This can be used for judging the validity of sPCA, as formalised by:

Proposition 1 (Structural Validity Condition)**.**

Let:

1. ρ *the constraint from Algorithm 1*
2. *Significant loadings exceed a threshold* $\tau \in (0, \rho]$*, defining a* confidence threshold *which defines the filtered support of each component:*

$$supp_\tau(sPC_k) := \{j \mid |(sPC_k)_j| \geq \tau\}, \quad s_k^\tau := |supp_\tau(sPC_k)|.$$

3. $\Delta_{\max}^\tau := \max_{a \neq b \leq K'} |supp_\tau(sPC_a) \cap supp_\tau(sPC_b)|, \quad s_{\max}^\tau := \max_{k \leq K'} s_k^\tau$

Suppose empirically: $\frac{\Delta_{\max}^\tau}{s_{\max}^\tau} \leq \delta$ *for some small* $\delta > 0$

Then: $sPC_1, \ldots, sPC_{K'}$ *approximate partitioned archetypes* $m_k^*(\theta_i)$ *at significance* τ*.*

Interpretation. This condition tests whether the leading sPCs exhibit low overlap (small δ), validating their interpretation as latent business types. The threshold τ sets the significance level within the budget defined by ρ in Algorithm 1, by including only loadings $\geq \tau$ in structural support testing. This empirically evaluates whether Lemma 1 holds. A large δ suggests the data does not support a modular interpretation.

sPCA Result. Table 1 reports Proposition 1 for different subsets; with $\rho = 2.414574$, the threshold $\tau = 0.1207287$ defines a 95% quasi-confidence interval. δ_4 and δ_5 are identical, with a sharp increase beyond $K' > 5$, narrowing valid choices to $K' \in \{4, 5\}$[6]. The overlap $\frac{\Delta_{\max}^\tau}{s_{\max}^\tau} \leq 0.1667$ falls within the accepted[7] threshold. The two subsets respectively explain 47.7% and 52.7% of the variance[8], respectively—consistent with Fact 1, which suggests a few dominant businesses shape the market. By Lemma 2, sPC1–5 represent core archetypes, while sPC6–20 capture refinements or noise.

[6] Only $K' > 3$ ensures all capacities are included.

[7] See Zou et al. (2006); Witten et al. (2009); Vu et al. (2013) for < 0.2 as a typical upper bound.

[8] By the definition of eigenvalue as a scale of projection, the proportion of explained variance by the ith component is $\frac{\lambda_i}{\sum_{j=1}^K \lambda_j}$, where λ_i is its eigenvalue.

K' leading sPC	4	5	6	7
$\delta_{K'}$	0.1667	0.1667	0.3334	0.5
Explained Var.	0.4766	0.5266	0.5830	0.6268

Table 1: Overlap ratio and Cumulative Proportion of Variance Explained by each K'

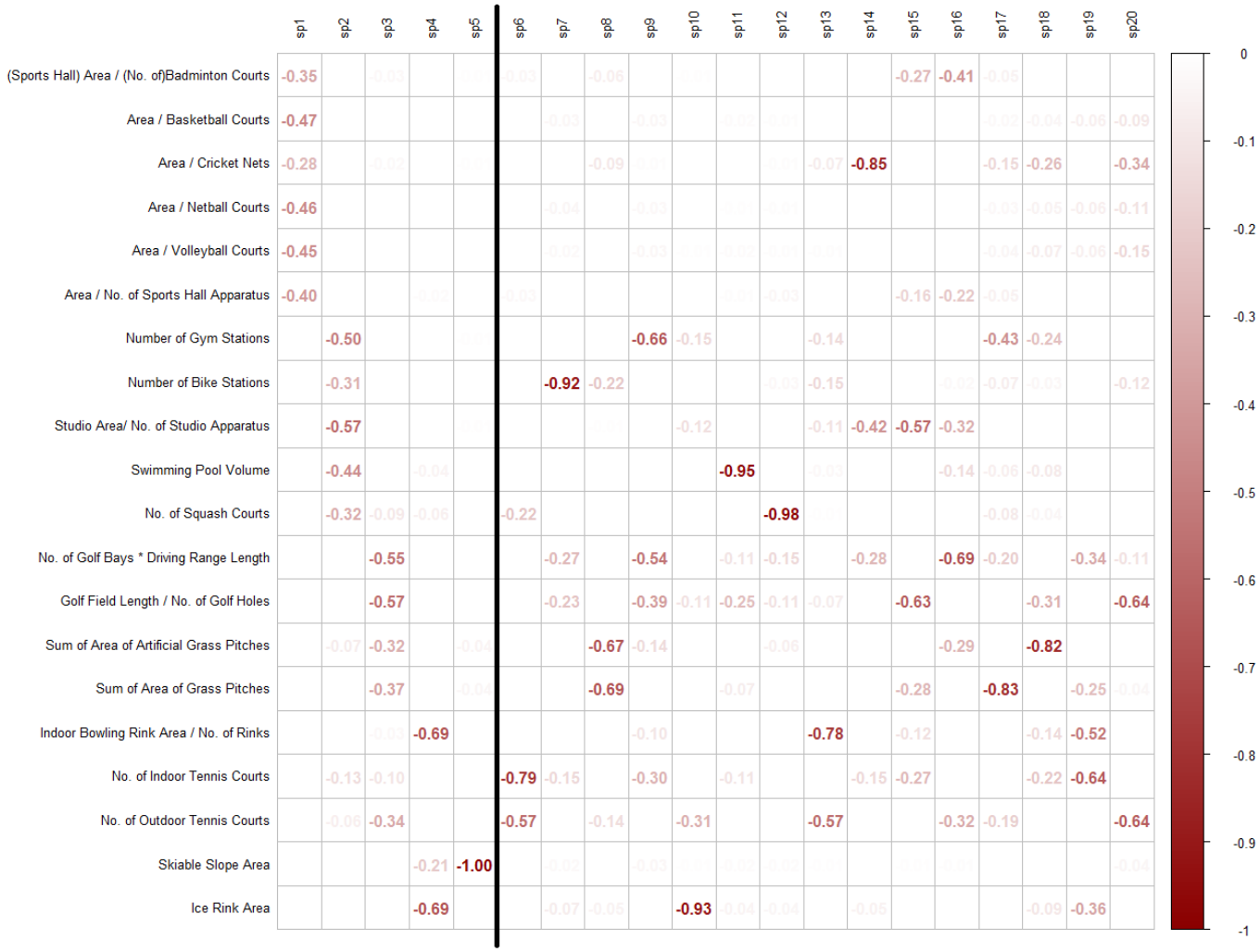

Figure 2: Estimated Principal Component Loadings

Figure 2 shows a clear partitioning of capacity types: sPC1 loads on sports halls (e.g., badminton, netball), sPC2 on fitness-oriented facilities (e.g., gyms, studios), sPC3 on golf and outdoor sports parks, while sPC4 and sPC5 capture specialised domains such as ice rinks and ski slopes. This **clustered, mostly disjoint** structure aligns with Lemma 1 and 2, where capacity bundles **map to distinct archetypes m_k^*, not arbitrary mixtures**.

3.2.3 Empirical Validation using LDA

Context. The sPCA analysis considered models with $K' \in \{4, 5\}$, noting potential overlap in ski-related capacities. LDA was used to: (1) validate whether each sPC reflects a distinct latent business configuration, and (2) assess whether lexical distinctions support the choice of $K' = 4$ or 5. If sPC-transformed sites are meaningful partitions of the market, clustering their names using topic modeling algorithm should yield corresponding partitions of words by conveying the similar facility combinations. Assumptions 5 and 6 enables LDA (Blei et al., 2003) - the algorithm applied to site names to identify semantic clusters in facility identity.

Algorithm 2 (LDA Topic Estimation with Weighted Contextual Tokens)**.**

1. ***Input****: Site names* $\{\texttt{name}_i\}$*, sPC scores* $\{sPC_{ik}\}$*, number of topics and words* $K = (k_t, k_w)$*.*
2. ***Preprocessing****:*
   - *Remove non-informative tokens (e.g., "Ltd", conjunctions).*
   - *Extract tokens and frequent neighbouring word pairs (window size* w*); define token set* T_i*.*
   - *Compute frequency of each neighbouring token pair across all sites and weight by frequency to refine token relevance.*
   - *Weight tokens by* $\eta(sPC_{ik})$*; test log, sqrt, exp. decay, and linear functions.*
3. ***Topic Inference****:*
   - *Assign tokens and weighted neighbour pairs in* T_i *to topics, scaled by* $\eta(sPC_{ik}) \cdot freq(w)$*.*
   - *Update topic-token* $\phi_k(w)$ *and document-topic* $\theta_i(k)$ *distributions:*

$$\phi_k(w) \propto \sum_i \eta(sPC_{ik}) \cdot freq(w) \cdot \delta(w \in T_i)$$

   - *Prune topics and tokens (*$K = (50, 50)$ *to* $(3, 5)$*) for interpretability.*
4. ***Output****: Contextual topic-token distributions* $\{\phi_k\}$*, aligned with sPC clusters.*

Note*: LDA assumes entities encode latent structure through word inclusion—**Lexical Signalling of Latent Structure** and **Inclusion-Based Semantic Mapping** (Blei et al., 2003), corresponding to Assumptions 5 and 6, respectively. Incorporating neighbouring token frequency refines semantic clarity and topic coherence, as frequent co-occurrence improves interpretability of latent structure (Mikolov et al., 2013; Chuang et al., 2012). Lastly, LDA was performed in Python using **sklearn** module.*

To see the word clusters from Algorithm 2 corresponds to the numerical partitions from Algorithm 1, the below formalises the conditions under which the latent structures recovered by Algorithm 1 and Algorithm 2 are aligned.

Lemma 3 (sPCA-LDA Alignment)**.**

Let:

1. $\mathbf{X} = [\mathbf{x}_i] \in \mathbb{R}^{n \times N}$ *where* $\mathbf{x}_i = m^*(\theta_i) + \boldsymbol{\varepsilon}_i$ *is archetype with noise.*
2. $sPC_1, \ldots, sPC_K$ *from* $\mathbf{X}$*, and* $\phi_1, \ldots, \phi_K$ *be weighted LDA topics on site names,* $\{\texttt{name}_i\}$*, with weights given by monotonic function* $\eta(sPC_{ik})$*.*

If:

1. *Lemma 2 holds,*
2. *The LDA topics are semantically distinct:* $supp(\phi_a) \cap supp(\phi_b) \approx \emptyset$ *for* $a \neq b$*,*

Then: sPCA and weighted LDA recover approximately the same modular archetypes $m_k^*(\theta_i)$*.*

Lemma 3 shows that, if the facility identities from numerical clusters from sPCA matches with the lexical clusters, sPCA and LDA recover the same latent archetypes from capacity data and site names. Empirical application requires adjustment for local noise, formalised by:

Lemma 4 (Stability of sPC-LDA under Noise)**.**

Let:

1. $\mathbf{X}_{true}$ *be the true partitioned-sparse matrix,*
2. $\mathbf{X}_{obs} = \mathbf{X}_{true} + \mathbf{E}$ *be observed matrix with noise* $\|\mathbf{E}\|_F \leq \varepsilon$*,*
3. sPC_k^{obs} *be the from* $\mathbf{X}_{obs}$ *that defines sPC-based topic proportions* θ_i^{sPC} *and topic-word distributions* ϕ_k^{sPC}*.*

If:

1. *Assumption 5 holds,*
2. *Proposition 1 holds,*
3. η *is Lipschitz*[9]*,*

[9]Stability under the weighting in Algorithm 2.

Then: $\|\theta_i^{sPC} - \mathbb{E}[\theta_i \mid D]\|_1 \leq C_1\varepsilon, \quad \|\phi_k^{sPC} - \mathbb{E}[\phi_k \mid D]\|_1 \leq C_2\varepsilon$, *with topic-word supports* ***approximately disjoint*** *up to first-order perturbations.*

Lemma 4 confirms that the sPCA–LDA alignment remains robust under bounded noise, indicating it is not a fragile artefact. Together, Lemmas 3 and 4 establish the sPCA-LDA alignment condition:

Proposition 2 (sPCA-LDA Validation)**.** *Let* $sPC_1, \ldots, sPC_K$ *be sPCs estimated from capacity data* $\mathbf{X}$*, and site names* $\{\texttt{name}_i\}$ *used to estimate topics* $\phi_1, \ldots, \phi_K$ *in Algorithm 2.*

Suppose empirically: Lemma 2, 3, and 4 hold; and Lexical signals align with sPC supports, both reflecting the same latent archetypes $m_k^*(\theta_i)$*.*

Then: the alignment of sPCA components with LDA topics validates that the sPCs represent meaningful latent business configurations, supported by both capacity data and semantic site names.

Interpretation. Distinct numerical and lexical modularity, plus word-number alignment, confirm that sPCs reflect coherent business types.

Table 2: sPCA-LDA Alignment of sPC1–sPC5

sPC	**Numerical Interpretation**	**Topic Words**	**Neighbouring Tokens**	**LDA-Validated Business Profile**
sPC1	Sports halls (badminton, netball)	*school*, *academy*, *college*	*community*, *sports*, *indoor*	Community sports halls; school or local authority venues
sPC2	Gym, studio, pool, bike stations	*fitness*, *leisure*, *centre*	*training*, *active*, *gym*	General fitness centres
sPC3	Golf, racquet courts, outdoor pitches	*golf*, *club*, *park*	*course*, *driving*, *fairway*	Golf and outdoor sports parks
sPC4	Ice rink, bowling rinks	*ice*, *rink*, *arena*, *planet*	*skate*, *cool*, *bowling*	Destination leisure venues (e.g., The Dome)
sPC5	Ski slopes, snow facilities	*ski*, *slope*, *snowboard*, *snowsports*	*dome*, *indoor*, *chill*, *alpine*	Snow centres (e.g., Snozone)

Note: Only representative topic words are presented.

LDA Result. Though not presented, topic–word distributions were stable across all weighting functions for both primary tokens and contextual neighbours, confirming the robustness of Assumptions 5 and 6. Table 2 shows clear alignment between numerical and lexical interpretations, supporting Proposition 2. Each sPC aligns with a distinct lexical cluster, supporting their modular archetype interpretation.

The selected sPC1-5 align with recognised facility types in FPM (Sport England, n.d.-b). sPC1–5 correspond to established business models: community sports halls, fitness centres, golf clubs and outdoor facilities, destination leisure venues, and snow sports facilities like Snozone, respectively. In contrast, sPC6–20 lacked interpretability or institutional relevance.

Therefore, sPC1–sPC5 were selected based on: (i) partitioned loadings (Proposition 1), (ii) semantic coherence (Proposition 2), and (iii) institutional alignment (FPM), as distinct, meaningful configurations.

3.3 Identification Strategy

Standard fixed effects (FE) estimators are vulnerable to endogeneity in this context due to several factors. Spatial spillovers arise when patients cross catchment boundaries or when regional characteristics—such as affluence—influence both health outcomes and infrastructure, undermining causal identification. Overlapping catchments mean a facility in one area can affect neighboring regions, biasing estimates. Additionally, neighborhood self-selection occurs when health-conscious individuals move to areas with better sports infrastructure, conflating pre-existing differences with treatment effects. Simultaneity also poses a challenge if infrastructure investments respond to an already active population. Lastly, omitted variable bias may result from broader, policy-driven shocks—such as concurrent infrastructure upgrades and public health campaigns linked to events like the London Olympics.

To address these concerns, lagged infrastructure values are employed as instruments. These are predetermined relative to current health outcomes and unaffected by contemporaneous selection or policy shocks, as infrastructure decisions typically precede implementation by several years. This lag structure helps mitigate simultaneity and restore orthogonality between treatment and unobservables.

The validity of this approach is supported both theoretically and empirically. Sports infrastructure exhibits path dependence and strong structural persistence due to its capital-intensive nature and planning frameworks (Baade & Dye, 1990; Wang et al., 2024). Under weak serial correlation, lagged supply decisions—formed without knowledge of current shocks—remain exogenous to contemporaneous errors. This makes lagged sPCs valid and relevant instruments in expectation-driven infrastructure markets. Accordingly, an instrumented fixed effects (IV-FE) specification—estimated via `xtivreg2` in STATA (Schaffer, 2010)—was adopted. Hausman tests favored IV-FE over random effects (Baltagi, 2005). Since the specification does not concern dynamic dependent variable, dynamic estimators such as Arellano-Bond (Arellano & Bond, 1991) were avoided to preserve interpretability and prevent distortion from further transformations (Athey & Imbens, 2017).

First-Stage (for $\mathbf{sPC}_k$).

$$\mathrm{sPC}_{it}^{(k)} = \sum_{j=1}^{5} \left(\pi_1^{(j)} \cdot \mathrm{sPC}_{i,t-1}^{(j)} + \pi_2^{(j)} \cdot \mathrm{sPC}_{i,t-2}^{(j)} \right) + \alpha^\top W_{it} + \mathrm{Time}_t + e_{it}, \tag{1}$$

for $k = 1, \ldots, 5$. Instruments Z_{it} are lags of all sPCs up to 2 lags,=. W_{it}: other covariates.

Joint instrumentalisation, grounded in rational expectations, addresses endogeneity by leveraging the dynamic nature of facility provision. Infrastructure development shows strong lagged dependencies, with persistence and inter-temporal substitution across archetypes (Romp & de Haan, 2007). Lagged sPCs capture

these market adjustments and gradual evolution while remaining exogenous to current shocks, mitigating simultaneity and reflecting structural persistence in capital-intensive developments.

Second-Stage.

$$Y_{it} = \sum_{k=1}^{5} \beta_k \cdot \widehat{\text{sPC}}_{it}^{(k)} + \delta \cdot \text{Integration}_{it}^{\text{site}} + \theta \cdot \text{StdHeat}_{it} + \gamma^\top X_{it} + \text{Time}_t + \epsilon_{it} \tag{2}$$

where Y_{it} is prevalence ratio in GP catchment i at time t, $\widehat{\text{sPC}}_{it}^{(k)}$ are first-stage fitted sPCs, $\text{Integration}_{it}^{\text{site}}$ and StdHeat_{it} are connectivity measures, and X_{it} are controls.

Clustered SE Standard errors were clustered to account for spatial correlation due to overlapping catchments and shared infrastructure. Following spatial econometrics and health geography literature (Diez Roux, 2004; Anselin & Rey, 2014; Cameron & Miller, 2015), clusters were defined using (1) area-weighted urban-rural classifications and (2) Primary Care Network (PCN) groupings within the most urban category. This two-tiered approach addresses spatial heterogeneity and local shocks while avoiding standard error underestimation (Bertrand et al., 2004; Abadie et al., 2023). Since over 25% of practices fell into the most urban level, additional stratification by PCN was necessary. Clusters ($G = 3{,}950$) were assigned as:

$$ID_{Cluster} = Urban + \mathbb{1}(Urban = \max\{Urban\}) \cdot PCN \times 10^{-7}$$

Here, *Urban* ranges from 0 to 11 (with up to 3 decimal points) and *PCN* from 1 to 1250 (integers), preserving urban classification while differentiating within highly urbanised areas.

Identification Assumptions

- **Relevance**: $\text{Cov}(Z_{it}, \text{sPC}_{it}^{(k)}) \neq 0$
- **Exogeneity under Weak Serial Correlation**: $\text{Cov}(Z_{it}, \epsilon_{it}) = 0$ for $Z_{it} \in \{\text{sPC}_{i,t-l}^{(k)}\}_{l \geq 2}$

Under the exogeneity under weak serial correlation, the lagged treatments are valid instruments, standard conduct in infrastructure-related panel data models as the assumptions guarantee the *asymptotic consistency* (Wooldridge, 2010). $T = 10$ is considered generally sufficient and well-suited (Baltagi, 2005; Wooldridge, 2010).

Despite their theoretical appeal, lagged instruments are not immune to validity concerns. **Dynamic simultaneity** poses a risk when health outcomes and infrastructure co-evolve over time, allowing lagged infrastructure to exert a **direct effect** on current health. In such settings, lagged variables may fail to satisfy the exclusion restriction if they are themselves outcomes of prior unobserved shocks that persist over

time. Bellemare and Wichman (2019) demonstrate that when residual direct effects from lagged regressors to the dependent variable are present, using them as instruments may *increase* bias rather than attenuate it. A related concern is the **long-term feedback loop**, whereby unobserved demand shocks—if serially correlated—can influence both infrastructure investments and health outcomes across multiple periods. This persistence can undermine instrument exogeneity and inflate the likelihood of Type I errors, effectively violating the core IV assumptions. Together, these dynamic interactions raise the risk of exclusion-restriction failure in settings where past infrastructure may not be fully orthogonal to current unobservables.

To test the robustness of the identification strategy, the model was subjected to a series of targeted stress tests:

- **Lead-placebo test:** The endogenous regressor sPC_t was replaced with its future value sPC_{t+1} to test for dynamic simultaneity. All $t+1$ coefficients were statistically insignificant, consistent with the timing assumption that future infrastructure capacity is not influenced by current or past health shocks ($t < t+1$).
- **Placebo outcome test:** The model was re-estimated using outcomes unrelated to physical activity, specifically the prevalence of epilepsy and chronic mental illness (e.g., schizophrenia). The lack of significant coefficients suggests that the instruments do not spuriously predict unrelated health outcomes, reinforcing the contextual exclusion restriction.
- **Near-outcome falsification:** Substituting the contemporaneous outcome Y_t with Y_{t+1} tested for long-run feedback effects. Estimated coefficients declined by approximately 20%, showing no sign of amplification over time and indicating the absence of latent long-run reverse causality.

Post-Estimation Diagnostics First-stage results confirm instrument relevance across all endogenous regressors, exceeding conventional Stock-Yogo thresholds for weak identification. The Kleibergen-Paap Wald F-statistic surpasses the 10% maximal IV relative bias threshold ($11.97 > 11.49$), supporting instrument strength despite non-i.i.d. errors. Underidentification is rejected (Kleibergen-Paap LM $\chi^2(6) = 31.45$, $p < 0.001$), indicating the instruments are sufficiently correlated with the endogenous regressors. The Hansen J statistic ($\chi^2(5) = 3.24$, $p = 0.66$) overidentification test fails to reject the null of valid instruments. Weak-instrument robust inference via Anderson-Rubin and Stock-Wright tests suggests joint insignificance cannot be rejected at conventional levels ($p \approx 0.13$). Overall, the results confirm strong identification and valid instruments.

Table 3: Description of Variables

Variable	Description
Dependent Variables	
Obesity	The proportion of registered patients aged 18 years or over with a BMI ≥ 30 in the preceding 12 months.
Sport Infrastructure	
SportHall (sPC1) Fitness (sPC2) Golf/Outdoor (sPC3) DLV (sPC4) Snow (sPC5)	Active Capacity within Catchment, normalised by the number of registered patients then projected onto the identified sPC. The qualitative profile follows from LDA result.
Area/Site	Catchment area covered by site within catchment on average.
Integration	Measure of integration.
Other Covariates	
BUA	Within-catchment BUA area (0.01km^2), per patient registered.
GreenWood	Area of greenspace (km^2, excluding areas included as part of grass pitch facilities) and woodland, per patient registered.
Never	Proportion of patients who haven't had an appointment since being registered with my current GP practice
GPconf	Proportion of patients who, during last general practice appointment, had confidence and trust in healthcare professional - Summary result - Yes (Combined 'yes, definitely' and 'yes, to some extent' responses, base excluding 'don't know / can't say')
Healthconf	Proportion of patients who answered as 'confident' in managing any issues arising from their condition (or conditions) - Summary result - Confident (Combined 'very confident' and 'fairly confident' responses)
NoSmok	Proportion of patient have never smoked
White	Proportion of ethnic group - White - English, Welsh, Scottish, Northern Irish or British
Asian	Proportion of ethnic group - Asian or Asian British - Any other Asian background (excl. Pakistani, Bangladeshi and Chinese)
Black	Proportion of ethnic group - Black, Black British, Caribbean or African - Any other Black, Black British, Caribbean or African background
Work	Proportion of patients with working status - In full-time paid work (30 hours or more each week) and part-time (under 30 hours)
Student	Proportion of patients with working status - In full-time education at school, college or university
Unemployed	Proportion of patients with working status - Unemployed
Pat.Ratio	$\frac{\text{Average No. of Weighted Patients}}{\text{Average No. of Registered Patients}}$ This captures the practice-level workload concentration. Higher value implies there are more clinical demand.
Payment	Average amount paid per weighted patient; payment adjusted by weighting based on the Carr-Hill formula (Carr-Hill et al., 1994) to account for patient workload at GP practice.
No. Pat	Total number of registered patients at GP practice, fiscal year from Apr 1st to Mar 31st.
Male	Of the total registered patients, proportion of male patients.
M:0-19	Proportion of males aged between 0 and 19.
M:20-49	between 20 and 49.
M:50-69	between 50 and 69.
F:0-19	Proportion of females aged between 0 and 19.
F:20-49	between 20 and 49.
F:50-69	between 50 and 69.

Lagged Infrastructure in First Stage

	SportHall	Fitness	Golf/Outdoor	DLV	Snow
L.SportHall	-0.0033	**-0.0139****	**-0.0141*****	**0.0669****	**0.0556*****
	(0.0051)	(0.0057)	(0.0042)	(0.0332)	(0.0182)
L.Fitness	**-0.0108****	0.0026	**0.0132*****	**-0.0613****	**-0.0350*****
	(0.0053)	(0.0044)	(0.0028)	(0.0297)	(0.0125)
L.Golf/Out	**0.0099***	**0.0220*****	**0.0062***	-0.0155	**-0.0327****
	(0.0052)	(0.0057)	(0.0034)	(0.0484)	(0.0157)
L.DLV	**-0.0057***	**-0.0147*****	**-0.0075*****	**0.4185*****	**0.1697*****
	(0.0031)	(0.0033)	(0.0021)	(0.0242)	(0.0561)
L.Snow	-0.0012	**0.0028*****	**0.0013****	**-0.1805*****	**-0.5030*****
	(0.0020)	(0.0007)	(0.0005)	(0.0153)	(0.0555)
L2.SportHall	**-0.0232*****	**-0.0235*****	**-0.0108*****	-0.0026	0.0088
	(0.0057)	(0.0055)	(0.0026)	(0.0155)	(0.0175)
L2.Fitness	**0.0122*****	**0.0119*****	**0.0079****	0.0185	**-0.0276*****
	(0.0041)	(0.0041)	(0.0034)	(0.0155)	(0.0102)
L2.Golf/Outdoor	-0.0033	-0.0024	**-0.0170*****	**-0.0644*****	0.0202
	(0.0048)	(0.0060)	(0.0035)	(0.0195)	(0.0221)
L2.DLV	**0.0117*****	**0.0279*****	**0.0151*****	**0.0998*****	**-0.0492***
	(0.0029)	(0.0058)	(0.0031)	(0.0151)	(0.0298)
L2.Snow	-0.0012	**-0.0038***	**-0.0020***	**0.0800*****	**0.3990*****
	(0.0012)	(0.0020)	(0.0012)	(0.0235)	(0.0478)

Summary Coefficient Table for Second Stage

	SportHall	Fitness	Golf/Outdoor	DLV	Snow
Coefficient	-0.0326	**0.0887****	-0.0420	**-0.0045***	0.0006
Std. Error	(0.0353)	(0.0408)	(0.0353)	(0.0023)	(0.0007)

Table 4: Lagged Regressions and Summary Results

Table 5: Table with Coefficients and Standard Errors (0 if value less than 10^6)

2nd Stage	From 2015-16			1st Stage		
0.0027* **	2016–17	**0.0035*** **	**0.0021*** **	**0.0012*** **	-0.0061	-0.0025
(0.0004)		(0.0005)	(0.0005)	(0.0003)	(0.0064)	(0.0028)
0.0043* **	2017–18	**-0.0135*** **	**-0.0216*** **	**-0.0141*** **	**-0.0141** **	0.0048
(0.0009)		(0.0008)	(0.0012)	(0.0006)	(0.0061)	(0.0038)
0.0075* **	2018–19	**0.0204*** **	**0.0126*** **	**0.0062*** **	**0.0144** **	**0.0363*** **
(0.0009)		(0.0009)	(0.0010)	(0.0006)	(0.0058)	(0.0061)
0.0116* **	2019–20	**-0.0209*** **	**-0.0108*** **	**-0.0100*** **	**-0.0266*** **	**0.0535*** **
(0.0011)		(0.0010)	(0.0011)	(0.0007)	(0.0066)	(0.0101)
-0.0261* **	2020–21	-0.0011	**0.0041*** **	**0.0014***	-0.0066	**-0.0225*** **
(0.0009)		(0.0009)	(0.0011)	(0.0008)	(0.0063)	(0.0053)
0.0054* **	2021–22	**0.0094*** **	**0.0042*** **	**0.0024*** **	**0.0147** **	**0.0085***
(0.0008)		(0.0009)	(0.0012)	(0.0009)	(0.0064)	(0.0048)
0.0226* **	2022–23	**-0.0040*** **	**-0.0029** **	-0.0010	-0.0089	-0.0068
(0.0008)		(0.0010)	(0.0013)	(0.0009)	(0.0068)	(0.0058)
-0.0001	Area/Site	0.0001	**0.0001** **	0.0001	0	0.0003
(0.0001)		(0.0000)	(0.0000)	(0.0000)	(0.0001)	(0.0003)
-0.0011	Integration	**0.0074** **	**0.0088***	**0.0052***	**0.0163** **	0.0014
(0.0023)		(0.0031)	(0.0045)	(0.0032)	(0.0076)	(0.0073)
0.1164	BUA	**-2.2036*** **	**-3.1406*** **	**-2.2243*** **	**-3.0135** **	-1.5601
(0.1267)		(0.4054)	(0.7020)	(0.4577)	(1.4090)	(1.4365)
0.367	GreenWood	**-4.1232*** **	**-2.7325*** **	**-3.8328*** **	1.0654	2.6404
(0.6208)		(1.1261)	(0.8629)	(0.8811)	(3.0293)	(3.3071)
-0.0101 **	Never	0.0021	0.0044	0.0038	-0.0184	0.0848
(0.0044)		(0.0063)	(0.0085)	(0.0050)	(0.0308)	(0.0807)
0.0209* **	Gpconf	**-0.0091** **	-0.0057	-0.0035	0.0032	**-0.0775** **
(0.0033)		(0.0045)	(0.0042)	(0.0027)	(0.0190)	(0.0336)
0.0077* **	HealthConf	0.0008	0.0049	0.0020	0.0106	0.0065
(0.0017)		(0.0032)	(0.0043)	(0.0019)	(0.0111)	(0.0218)
-0.0033*	NoSmok	0.0016	0.0019	0.0015	0.0088	0.0069
(0.0017)		(0.0024)	(0.0029)	(0.0019)	(0.0099)	(0.0162)
-0.0023	White	-0.0033	-0.0035	-0.0017	-0.0253	0.0303
(0.0023)		(0.0037)	(0.0054)	(0.0035)	(0.0249)	(0.0272)
0.0028	Asian	0.0078	0.0080	0.0058	0.0405	0.0571
(0.0046)		(0.0105)	(0.0126)	(0.0084)	(0.0370)	(0.0487)
0.0015	Black	-0.0036	**-0.0245** **	**-0.0118** **	-0.0502	-0.0301
(0.0057)		(0.0094)	(0.0095)	(0.0057)	(0.0388)	(0.0370)
-0.0071* **	Work	0.0043	0.0069	**0.0061** **	**-0.0258***	-0.0367
(0.0023)		(0.0035)	(0.0049)	(0.0025)	(0.0155)	(0.0334)
-0.0074*	Student	-0.0056	-0.0069	-0.0014	-0.0133	-0.0319
(0.0043)		(0.0057)	(0.0076)	(0.0043)	(0.0267)	(0.0557)
0.005	Unemployed	-0.0013	0.0018	0.0026	0.0376	0.0900
(0.0042)		(0.0076)	(0.0050)	(0.0031)	(0.0270)	(0.0592)
0.0547* **	Pat.Ratio	-0.0076	0.0169	-0.0029	0.1057	-0.0548
(0.0113)		(0.0098)	(0.0169)	(0.0102)	(0.0728)	(0.0991)
0.0000* **	Payment	0	0	0	-0.0001	0
(0)		(0.0000)	(0.0000)	(0.0000)	(0.0001)	(0.0000)
-0.0000* **	No. Pat	0	0	0	0	**0.0000***
(0)		(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
0.1079	Male	-0.0629	0.2341	0.1130	-0.0754	-1.8120
(0.1247)		(0.1264)	(0.1636)	(0.1140)	(0.6331)	(1.1500)
0.0115	M:0–19	0.0141	-0.1285	-0.0327	0.4198	1.1092
(0.0846)		(0.1008)	(0.1619)	(0.0933)	(0.4897)	(0.9369)
-0.1367*	M:20–49	0.0370	-0.0176	-0.0015	0.2527	0.3236
(0.0735)		(0.0848)	(0.1128)	(0.0705)	(0.3890)	(0.8074)
-0.1504 **	M:50–69	0.0688	**0.2446** **	**0.1451** **	0.1534	0.5717
(0.0755)		(0.0630)	(0.1012)	(0.0628)	(0.3939)	(0.6547)
0.1295*	F:0–19	-0.0258	0.0529	0.0380	0.4465	-1.0607
(0.0712)		(0.0696)	(0.0801)	(0.0588)	(0.3672)	(0.7039)
0.0304	F:20–49	-0.0312	**0.2451** **	0.0853	0.1975	**-1.0897** **
(0.0614)		(0.0592)	(0.1050)	(0.0694)	(0.3367)	(0.5200)
0.2658* **	F:50–69	-0.0881	-0.0863	-0.0717	0.2361	**-1.3292***
(0.0742)		(0.0607)	(0.0791)	(0.0564)	(0.3670)	(0.7076)

4 Result

Note: Negative sPC loadings imply **higher facility capacity supply**.

4.1 First-Stage

Some notable points emerged from the first-stage regression. Lagged archetype variables (L, $L2$ for $t-1$, $t-2$; Table **??**) were generally significant, indicating lagged complementarity (positive coeff.) and substitutability (negative) among businesses. Diagonal coefficients capture prior supply effects on the same facility type, while off-diagonal capture the effects from the other facilities. For $t-2$, all businesses were highly significant in the current realisation, indicative of weak-serial correlation. There were altering directions across $t-1$ and $t-2$, supporting the business dynamism as expected. For instance, `DLV` showed strong, persistent effect yet its direction altered as time elapsed while `Fitness` was complementary to `Golf/Outdoor` across the times. Since `Fitness` is indoor facility, it would be complementary to outdoor facilities. As such, results reveal **inter-temporal dependencies** across businesses.

Importantly, `Integration` emerged as key latent factor in the first-stage regression. The significant positive signs except for `Snow` suggests that **better overall connectivity** between the residential areas and sites was **negatively associated** with the contemporaneous supply of infrastructure. This implies saturation or strategic substitution in connected areas, echoed by urban spatial dynamics (Batty, 2009). Given the insignificance of `Area/Site` the results highlight that **spatial accessibility shapes supply-demand dynamics**.

`BUA` and `GreenWood` were positively linked to infrastructure provision, indicating urban concentration and highlighting equity concerns for rural areas (Woods, 2005) and greenspace reliance (Peters et al., 2010).

Although no consistent trend emerged, but notable fluctuations in facility supply occurred around 2018–20, coinciding with the COVID-19 pandemic. A decline in 2018–19 reflects lockdown closures (fiscal year: April–March). Supply rebounded in 2019–20, suggesting disruptions were largely temporary, consistent with national data on phased reopenings (Sport England, 2021; DCMS, 2021).

The remaining characteristic variables were largely insignificant. Once controlled for lagged infrastructure which are forward-looking optimised existing stock, regional characteristics had virtually no impact on the contemporaneous supply, consistent with the rational expectation outcome. This supports the use of lagged variable as IVs.

4.2 Second-Stage

Time Effects Time trends (Table 5) show a general rise in recorded prevalence, indicating an *exacerbating trend* in physical health requiring policy attention. Obesity showed a sharp drop in 2020-21, likely due to disrupted BMI recording during COVID-19 (Robinson et al., 2021). In 2022-23, it showed a major spike around five times higher than 2021-22, exceeding national post-pandemic trends (NHS Digital, 2024b).

The model captures *localised time trends* by adjusting for infrastructure, demographics, and socio-economic factors, suggesting national averages may obscure deeper disparities. A likely explanation is widening post-pandemic inequalities: obesity rates stabilised in affluent areas but rose in deprived ones (Darmon & Drewnowski, 2008). This is supported by the significance of `Payment` for obesity, indicating higher burdens in poorer catchments.

Adjusting for spatial integration and clustering, the model uncovers disparities absent in national data. Time effects align with national trends, confirming robustness. This spike is not an artefact of estimation—it reflects a real, disproportionate rise in obesity across deprived catchments, possibly exacerbated by post-pandemic inequities.

Environmental and Spatial Integration The insignificance of environmental and spatial integration variables reflects their **indirect influence** in the IV-FE model. As sports infrastructure was instrumented via lagged covariates, environmental effects were likely absorbed in the first stage, influencing **infrastructure provision rather than health outcomes directly**. While factors like `GreenWood` and connectivity shape *facility distribution*, they show no direct effect after adjusting for infrastructure, socio-economic, and temporal variables. This suggests environmental effects influence access and supply, **not direct** health prevalence.

Demographic Results reveal age and gender disparities in health. Compared to the 70+ baseline, ageing was positively associated with male obesity, while `F:50-69` showed a negative link - likely due to hormonal factors like menopause and testosterone decline (Lovejoy & Sainsbury, 2009).

Sport Infrastructure `Fitness` was significantly associated with lower prevalence, while `DLV` showed a marginal positive link (Table **??**). Fitness centres support structured, moderate-to-vigorous physical activity effective in reducing BMI and preventing obesity (Jakicic et al., 2018), including resistance training and cardio classes. In contrast, DLVs offer light activity like bowling with limited weight-loss effect (Warburton et al., 2006).

A one σ increase in `Fitness` corresponded to $\approx 9\%$ reduction in obesity prevalence, while `DLV` showed a

0.5% increase, likely due to contextual factors. DLVs near malls and food courts may offset activity benefits due to calorie access (Fraser et al., 2010). Social habits like snacking or drinking during leisure may further reduce health gains.

5 Discussion

5.1 Policy Implications

Plan Strategically: Health outcomes depend on facility *type*, *capacity*, and *location*. *Data-driven investment* should align with local demand and health needs. However, spatial integration did not have a direct influence on health outcomes, implying that well-connected areas may need fewer new facilities. Policymakers should focus on expanding infrastructure in underserved areas to improve access equity.

Prioritise High-Impact Facilities: Focus on *gyms*, which significantly reduce obesity, respectively. As their provision often falls under local discretion, a more top-down approach may be warranted. Funding from initiatives like the Multisport Grassroots Programme (HM Government, 2023) could be redirected from pitch-based to higher-impact venues.

Advance Equity: Balance investments across *deprived* areas, prioritising high-need communities. The 2022–23 obesity spike, likely driven by economic hardship, underscores the need for targeted action. The strong link between *deprivation* and obesity (0.3% increase per £100 in `Payment`) suggests that investing in sports infrastructure in deprived areas may ease GP burdens and prevent further health decline.

5.2 Limitations

Although post-estimation diagnostics confirmed instrument robustness, validity holds **only if** the relevance and exogeneity assumptions are met. While relevance is defensible, exogeneity—though fundamentally untestable—is justified by rational expectations, a potentially fragile assumption. This paper regrettably admit that, though it sought for stress-tests, the framework should still undergo other robustness tests - ideally with guaranteed exogenous data. sPCA assumes site-capacity partial equilibrium, which may be unstable due to COVID-19 shocks. Therefore the study remains observational, while joint instrumentalisation better addresses covariate interdependence than standard fixed effects models.

An exogeneity test using Sport England's grant data (Sport England, n.d.-c) showed stable results, with Golf/Outdoor becoming insignificant—supporting its likely spurious effect. Grant data were excluded from the main model due to ambiguity over whether funding expanded capacity, raising concerns about measurement error and artificial endogeneity. More detailed grant data could strengthen this and related research on infrastructure provision.

Future research could improve robustness using propensity score matching (PSM), aligning treated and control groups by covariates (Rosenbaum & Rubin, 1983; Stuart, 2010; Imbens & Rubin, 2015). Continuous

treatment PSM (Hirano & Imbens, 2004) offers a viable approach, though generating valid propensity scores from multiple covariates remains challenging.[10]

Alternatively, linking infrastructure data to English Prescribing Data (NHS Business Services Authority, 2024) could refine outcome measures. Prescriptions for weight-loss drugs like orlistat serve as proxies for obesity, enabling finer temporal analyses. However, infrastructure evolves slowing, complicating alignment with dynamic clinical indicators.[11]

6 Concluding Remark

This study presents a comprehensive framework to assess how the structure, density, and spatial integration of sports infrastructure shape population-level health outcomes—specifically obesity. Moving beyond simple facility counts, it integrates sPCA, LDA, and heat diffusion modelling to capture latent heterogeneity in provision, including co-location, capacity depth, and real-world accessibility. IV-FE regressions enhance causal inference by addressing endogeneity and multicollinearity.

The contributions are threefold. First, the study validates latent business archetypes by aligning sPCA-derived profiles with LDA-based semantic analysis—offering a novel method for analysing complex facility data. Second, it identifies heterogeneous health effects: gyms reduce obesity, and leisure venues show limited impact, underscoring the need to target high-impact infrastructure. Third, it reveals spatial and socio-economic disparities, including a catchment-level obesity surge in 2022–23, likely driven by post-pandemic inequalities and obscured in national trends.

These findings support more strategic, data-driven, and equitable infrastructure planning. Policymakers should prioritise facilities with demonstrated health impacts, close access gaps in deprived areas, and integrate infrastructure into public health strategies. While observational, the study lays a foundation for future research using granular datasets like APD and encourages methodological innovation in public health and urban planning.

[10] This study conducted PSM and the results were broadly consistent but sensitive to caliper choices, limiting robustness.

[11] This study explored staggered rollouts and policy pilots as natural experiments but faced limited short-term variability. Future work could develop time-series models using APD facility open/close data to improve causal inference.

7 Appendix

7.1 sPCA-LDA Assumptions

Assumption 1 is supported by the decade-long panel ($T = 10$), allowing agents to learn (Sargent, 1993; Evans & Honkapohja, 2001). Rational expectations are standard for moderate T (Blundell & Bond, 1998; Arellano & Bover, 1995), especially in capital-intensive markets (Muth, 1961; Romp & de Haan, 2007).

Assumption 2 arises from constrained choice: agents select optimal capacity bundles (Varian, 1992; Mas-Colell et al., 1995), with zeros reflecting non-selection (Eaton & Schmitt, 1994; Neary, 2003). As in **Pre-treatment**, N/A values represent unchosen, not unavailable, capacities.

Assumption 3 reflects real-world constraints on feasible business types, aligning with finite technology (McFadden, 1981) and discrete choice models (Berry, 1994; Train, 2009), where market forces yield a few efficient configurations (Sutton, 1991; Bresnahan & Reiss, 1991).

Assumption 4 follows from segmented preferences (Fact 2) and assumes a finite number of consumer types (McFadden, 1981; Anderson et al., 1992; Train, 2009), consistent with urban sorting models (Becker & Murphy, 2000; Bayer et al., 2004).

Assumption 5 draws from branding theory: firm names signal service type and market position (Pohlmann & Opitz, 2013; Klink, 2003), with common terms like "fitness" or "lido" conveying industry norms.

Assumption 6 reflects that word presence conveys meaning (Blei et al., 2003; Manning et al., 2008); e.g., "gym and fitness" $\equiv$ "fitness and gym".

7.2 Sketch Proofs

Lemma 1. By KKT conditions, a group G is active if

$$\sum_{i \in G_k} (\mathbb{E}[p_i] - c'(x_i))x_i \geq S_{G_k}$$

for some $x_i > 0$. Thus, $x_j = 0$ for $j \notin \bigcup_{k \in K_i^*} G_k$, and $\mathbf{x}_i^*$ is a convex combination of finitely many sparse archetypes $m_1, \ldots, m_K$. □

Lemma 2. $\mathbf{X}$ has sparse columns. Sparse PCA extracts components with limited nonzeros. Leading K' components align with disjoint supports (δ small), while $k > K'$ components capture noise and overlap. □

Proposition 1. With L_1 sparsity and threshold τ, if

$$\Delta_{\max}^{\tau} / s_{\max}^{\tau} \leq \delta,$$

then the first K' sPCs load mainly on distinct supports. □

Lemma 3. Given $\mathbf{x}_i = m^*(\theta_i) + \varepsilon_i$:

- sPC_1–$\text{sPC}_{K'}$ approximate $m_k^*(\theta_i)$,
- LDA weights $\eta(\text{sPC}_{ik})$ align topics with archetypes,
- Disjoint vocabularies imply cluster alignment.

Thus, sPCA and LDA recover the same latent structure. □

Lemma 4. With $\mathbf{X}_{\text{obs}} = \mathbf{X}_{\text{true}} + E$ and $\|E\|_F \leq \varepsilon$:

- $\|\text{sPC}_k^{\text{obs}} - \text{sPC}_k^{\text{true}}\|_1 = O(\varepsilon)$,
- η perturbations are $O(\varepsilon)$,
- Topic proportions θ_i^{sPC} deviate by $O(\varepsilon)$.

Thus, topic recovery remains stable under bounded noise. □

Proposition 2. By Lemmas 2–4, sPCs and LDA topics consistently reflect the same archetypes, validating the sPCs as meaningful configurations. □

7.3 Why Uncentered sPCA?

Although centering is standard in PCA, uncentred PCA is commonly used when data are naturally non-negative or defined on $\mathbb{R}_0^+$ (van den Dool, 2007; Braak, 1983; Jackson, 1991; Reyment and Jöreskog, 1993). Centering removes mean bias to highlight relative variation, but it discards the original scale—which can be critical when the origin carries real meaning (Brockwell, 1992).

In this setting, site-level capacity data are strictly non-negative and reflect actual allocations. By Assumption 2, each site optimises over a full capacity set, selecting a subset with positive entries and zeros for excluded types. Here, $\vec{0}$ is meaningful—it encodes strategic non-selection, as formalised in Lemma 1.

PCA leverages covariance to uncover such structure: positive co-occurrence implies complementarity (e.g., gym + studio), while negative correlation suggests substitutability. These patterns reflect constrained revealed preferences. Since capacities are bounded below by zero, not all trade-offs are observable. For example, if negative values were feasible, one might reduce indoor offerings to expand outdoor provision—a counterfactual hidden by non-negativity. Centering makes all excluded types appear equally unpreferred, distorting economic meaning. In reality, zeros vary in intent—some are deliberately excluded, others merely infeasible.

Thus, centering distorts the economic meaning of zeros. Uncentered sPCA, by contrast, preserves the sparsity pattern and respects the partitioning structure essential for model recovery.

References

[1] Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? The Quarterly Journal of Economics, 138(1), 1–35. https://doi.org/10.1093/qje/qjac043

[2] Abel, G. A., Barclay, M. E., & Payne, R. A. (2013). Adjusted indices of multiple deprivation to enable comparisons within and between constituent countries of the UK including an illustration using mortality rates. BMJ Open, 3(6), e002750. https://doi.org/10.1136/bmjopen-2013-002750

[3] Albert, P. R. (2015). Why is depression more prevalent in women? Journal of Psychiatry & Neuroscience, 40(4), 219–221. https://doi.org/10.1503/jpn.150205

[4] Allender, S., Cowburn, G., & Foster, C. (2006). Understanding participation in sport and physical activity among children and adults: A review of qualitative studies. Health Education Research, 21(6), 826–835. https://doi.org/10.1093/her/cyl063

[5] An, R., & Sturm, R. (2015). School and residential neighborhood food environment and diet among California youth. American Journal of Preventive Medicine, 49(4), 529–535. https://doi.org/10.1016/j.amepre.2015.02.018

[6] Anderson, S. P., de Palma, A., & Thisse, J.-F. (1992). Discrete choice theory of product differentiation. MIT Press.

[7] Anselin, L., & Rey, S. J. (2014). Modern spatial econometrics in practice: A guide to GeoDa, GeoDaSpace and PySAL. GeoDa Press LLC.

[8] Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Review of Economic Studies, 58(2), 277–297. https://doi.org/10.2307/2297968

[9] Arellano, M., & Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. Journal of Econometrics, 68(1), 29–51. https://doi.org/10.1016/0304-4076(94)01642-D

[10] Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. Journal of Economic Perspectives, 31(2), 3–32. https://doi.org/10.1257/jep.31.2.3

[11] Baade, R. A., & Dye, R. F. (1990). The impact of stadium and professional sports on metropolitan area development. Growth and Change, 21(2), 1–14. https://doi.org/10.1111/j.1468-2257.1990.tb00513.x

[12] Baltagi, B. H. (2005). Econometric analysis of panel data (3rd ed.). John Wiley & Sons.

[13] Batty, M. (2009). Cities as complex systems: Scaling, interaction, networks, dynamics and urban morphologies. In R. A. Meyers (Ed.), Encyclopedia of Complexity and Systems Science (pp. 1041–1071). Springer. https://doi.org/10.1007/978-0-387-30440_69

[14] Bayer, P., McMillan, R., & Rueben, K. (2004). An equilibrium model of sorting in an urban housing market. National Bureau of Economic Research Working Paper No. 10865. https://doi.org/10.3386/w10865

[15] Becker, G. S., & Murphy, K. M. (2000). Social economics: Market behavior in a social environment. Harvard University Press.

[16] Berry, S. (1994). Estimating discrete-choice models of product differentiation. RAND Journal of Economics, 25(2), 242–262. https://doi.org/10.2307/2555829

[17] Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? The Quarterly Journal of Economics, 119(1), 249–275. https://doi.org/10.1162/003355304772839588

[18] Biddle, S. J. H., Ciaccioni, S., Thomas, G., & Vergeer, I. (2019). Physical activity and mental health in children and adolescents: An updated review of reviews and an analysis of causality. Psychology of Sport and Exercise, 42, 146–155. https://doi.org/10.1016/j.psychsport.2018.11.010

[19] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[20] Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. Journal of Econometrics, 87(1), 115–143. https://doi.org/10.1016/S0304-4076(98)00009-8

[21] Bresnahan, T. F., & Reiss, P. C. (1991). Entry and competition in concentrated markets. Journal of Political Economy, 99(5), 977–1009. https://doi.org/10.1086/261786

[22] Brockwell, P. J. (1992). Application of factor analysis to spectroscopic methods. [Doctoral Thesis, University of Greenwich]. University of Greenwich. https://gala.gre.ac.uk/id/eprint/6117/.

[23] Call, J. B., & Shafer, K. (2016). Gendered manifestations of depression and help seeking among men. American Journal of Men's Health, 10(1), 86–96. https://doi.org/10.1177/1557988315623993

[24] Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. Journal of Human Resources, 50(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

[25] Carr-Hill, R., Sheldon, T., Smith, P., Martin, S., Peacock, S., & Hardman, G. (1994). Allocating resources to general practice: Formula development. Centre for Health Economics, University of York.

[26] Catchment Based Approach. (n.d.). GP catchments based on patient registration patterns. https://catchmentbasedapproach.org/

[27] Chandra, A., & Skinner, J. (2012). Technology growth and expenditure growth in health care. Journal of Economic Literature, 50(3), 645–680. https://doi.org/10.1257/jel.50.3.645

[28] Choi, S., Kim, Y., & Lee, J. (2023). Gender differences in depression: The role of caregiving stressors and social support. Journal of Affective Disorders, 320, 123–130. https://doi.org/10.1016/j.jad.2023.01.045

[29] Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. In Proceedings of the International Working Conference on Advanced Visual Interfaces (pp. 74–77). ACM. https://doi.org/10.1145/2254556.2254572

[30] Coalter, F. (2013). Sport for development: What game are we playing? Routledge.

[31] Cutler, D. M., & McClellan, M. (2001). Is technological change in medicine worth it? Health Affairs, 20(5), 11–29. https://doi.org/10.1377/hlthaff.20.5.11

[32] Darmon, N., & Drewnowski, A. (2008). Does social class predict diet quality? The American Journal of Clinical Nutrition, 87(5), 1107–1117. https://doi.org/10.1093/ajcn/87.5.1107

[33] DCMS. (2021). COVID-19 Response - Spring 2021. https://www.gov.uk/government/publications/covid-19-response-spring-2021/covid-19-response-spring-2021

[34] Development Team. (2023). QGIS Geographic Information System [Software]. Open Source Geospatial Foundation. https://qgis.org

[35] Diez Roux, A. V. (2004). Estimating neighborhood health effects: The challenges of causal inference in a complex world. Social Science & Medicine, 58(10), 1953–1960. https://doi.org/10.1016/j.socscimed.2003.08.008

[36] Downward, P., & Rasciute, S. (2011). Does sport make you happy? An analysis of the well-being derived from sports participation. International Review of Applied Economics, 25(3), 331–348. https://doi.org/10.1080/02692171.2010.511168

[37] Downward, P., Hallmann, K., & Rasciute, S. (2024). Sport participation, health and wellbeing: A longitudinal analysis. International Journal of Sport Policy and Politics, 16(2), 231–250. https://doi.org/10.1080/19406940.2024.2404949

[38] Eaton, B. C., & Schmitt, N. (1994). Flexible manufacturing and market structure. American Economic Review, 84(4), 875–888.

[39] Eime, R. M., Young, J. A., Harvey, J. T., Charity, M. J., & Payne, W. R. (2013). A systematic review of the psychological and social benefits of participation in sport for adults: Informing development of a conceptual model of health through sport. International Journal of Behavioral Nutrition and Physical Activity, 10, 135. https://doi.org/10.1186/1479-5868-10-135

[40] Evans, G. W., & Honkapohja, S. (2001). Learning and expectations in macroeconomics. Princeton University Press.

[41] Fraser, L. K., Edwards, K. L., Cade, J., & Clarke, G. P. (2010). The geography of fast food outlets: A review. International Journal of Environmental Research and Public Health, 7(5), 2290–2308. https://doi.org/10.3390/ijerph7052290

[42] Guerra-Urzola, R., & Ramírez-Hassan, A. (2021). Bayesian treatment effects due to a subsidized health program: The case of preventive health care. Psychometrika, 86(3), 1–24. https://doi.org/10.1007/s11336-021-09773-2

[43] Gravelle, H., Sutton, M., & Ma, A. (2003). Doctor behaviour under a pay for performance contract: Treating, cheating and case finding? The Economic Journal, 113(485), 122–140. https://doi.org/10.1111/1468-0297.00097

[44] Green, B. C. (2008). Sport as an agent for social and personal change. In V. Girginov (Ed.), Management of Sports Development (pp. 129–145). Elsevier.

[45] Hallmann, K., Wicker, P., Breuer, C., & Schönherr, L. (2012). Understanding the importance of sport infrastructure for participation in different sports—Findings from multi-level modeling. European Sport Management Quarterly, 12(5), 525–544. https://doi.org/10.1080/16184742.2012.687756

[46] Higgs, P., & Gilleard, C. (2015). Key social and cultural drivers of changes affecting trends in attitudes and behaviour throughout the ageing process and what they mean for policymaking. Foresight, Government Office for Science. https://assets.publishing.service.gov.uk/media/5a75bd31e5274a436829999b/gs-15-14-future-ageing-attitudes-social-cultural-er05.pdf

[47] Hirano, K., & Imbens, G. W. (2005). The propensity score with continuous treatments. In A. Gelman & X.-L. Meng (Eds.), Applied Bayesian modeling and causal inference from incomplete-data perspectives (pp. 73–84). Wiley.

[48] HM Government (2022). Health inequalities dashboard: June 2022 data update. https://www.gov.uk/government/statistics/health-inequalities-dashboard-june-2022-data-update

[49] HM Government. (2023). Get Active: A strategy for the future of sport and physical activity. Department for Culture, Media and Sport. https://www.gov.uk/government/publications/get-active-a-strategy-for-the-future-of-sport-and-physical-activity

[50] Imbens, G. W., & Rubin, D. B. (2015). Causal inference for statistics, social, and biomedical sciences: An introduction. Cambridge University Press.

[51] Jackson, J. E. (1991). A user's guide to principal components. Wiley.

[52] Jakicic, J. M., Rogers, R. J., Davis, K. K., & Collins, K. A. (2018). Role of physical activity and exercise in treating patients with overweight and obesity. Clinical Chemistry, 64(1), 99–107. https://doi.org/10.1373/clinchem.2017.272443

[53] Kaczynski, A. T., & Henderson, K. A. (2007). Environmental correlates of physical activity: A review of evidence about parks and recreation. Journal of Physical Activity and Health, 4(4), 619–632. https://journals.humankinetics.com/view/journals/jpah/4/4/article-p619.xml

[54] Katz, L. (1953). A new status index derived from sociometric analysis. Psychometrika, 18(1), 39–43. https://doi.org/10.1007/BF02289026

[55] Klink, R. R. (2003). Creating meaningful brands: The relationship between brand name and brand mark. Marketing Letters, 14(3), 143–157. https://doi.org/10.1023/A:1027476132607

[56] Kokolakakis, T., Lera-López, F., & Castellanos, P. (2014). Regional differences in sports participation: The case of local authorities in England. International Journal of Sport Finance, 9(2), 149–172. https://shura.shu.ac.uk/10323/1/Kokolakakis.pdf

[57] Kokolakakis, T., Lera-López, F., & Castellanos, P. (2014). Regional differences in sports participation: The case of Local Authorities in England. International Journal of Sport Finance, 9(2), 149–172.

[58] Kokolakakis, T., Lera-López, F., & Panagouleas, T. (2012). Analysis of the determinants of sports participation in Spain and England. Applied Economics, 44(21), 2785–2798. https://doi.org/10.1080/00036846.2011.566204

[59] Kontopantelis, E., Mamas, M. A., van Marwijk, H., Ryan, A. M., & Doran, T. (2015). Chronic morbidity, deprivation and primary medical care spending in England in 2011/12: a cross-sectional spatial analysis. BMC Medicine, 13, 68. https://doi.org/10.1186/s12916-015-0305-4

[60] Langford, B. C., Cherry, C. R., Bassett, D. R., Jr., & Dhakal, N. (2017). Comparing physical activity of pedal-assist electric bikes with walking and conventional bicycles. Journal of Transport & Health, 6, 463–473. https://doi.org/10.1016/j.jth.2017.06.002

[61] Lovejoy, J. C., & Sainsbury, A. (2009). Sex differences in obesity and the regulation of energy homeostasis: Etiology and pathophysiology. Obesity Reviews, 10(2), 154–167. https://doi.org/10.1111/j.1467-789X.2008.00529.x

[62] Lubans, D. R., Richards, J., Hillman, C. H., Faulkner, G., Beauchamp, M. R., Nilsson, M., ... & Biddle, S. J. H. (2016). Physical activity for cognitive and mental health in youth: A systematic review of mechanisms. Pediatrics, 138(3), e20161642. https://doi.org/10.1542/peds.2016-1642

[63] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

[64] Marmot, M., Allen, J., Goldblatt, P., Boyce, T., McNeish, D., Grady, M., & Geddes, I. (2010). Fair society, healthy lives: The Marmot Review. Institute of Health Equity. https://www.instituteofhealthequity.org/resources-reports/fair-society-healthy-lives-the-marmot-review

[65] Marmot, M., Allen, J., Goldblatt, P., Herd, E., & Morrison, J. (2020). Build back fairer: The COVID-19 Marmot review. Institute of Health Equity. https://www.instituteofhealthequity.org/resources-reports/build-back-fairer-the-covid-19-marmot-review

[66] McFadden, D. (1981). Econometric models of probabilistic choice. In C. F. Manski & D. McFadden (Eds.), Structural analysis of discrete data with econometric applications (pp. 198–272). MIT Press.

[67] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26, 3111–3119. https://papers.nips.cc/paper_files/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

[68] Muth, J. F. (1961). Rational expectations and the theory of price movements. Econometrica, 29(3), 315–335. https://doi.org/10.2307/1909635

[69] Neary, J. P. (2003). Globalization and market structure. Journal of the European Economic Association, 1(2–3), 245–271. https://doi.org/10.1162/154247603322390928

[70] NHS Business Services Authority. (2024). English prescribing data: Detailed prescribing information by practice (2013–2023) [Data set]. https://opendata.nhsbsa.net/dataset/english-prescribing-data-epd

[71] NHS Digital. (2024a). Quality and Outcomes Framework (QOF), 2022–23. https://digital.nhs.uk/data-and-information/publications/statistical/quality-and-outcomes-framework-achievement-prevalence-and-exceptions-data/2022-23

[72] NHS Digital. (2024b). National Obesity Audit, July 2024 – September 2024. https://digital.nhs.uk/data-and-information/publications/statistical/national-obesity-audit/national-obesity-audit-july-2024---september-2024

[73] NHS Digital. (n.d.). Quality and Outcomes Framework (QOF)[Data set]. https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/quality-and-outcomes-framework-qof

[74] NHS Digital. (2013–2023a). Quality and Outcomes Framework (QOF) data, 2013–2023 [Data set]. https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/quality-and-outcomes-framework-qof

[75] NHS Digital. (2013–2023b). Patients registered at a GP practice [Data set]. https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice

[76] NHS England. (2013–2023). NHS payments to general practice, England [Data set]. https://www.england.nhs.uk/publication/nhs-payments-to-general-practice-england/

[77] Nichols, J., Giles-Corti, B., & Knuiman, M. (2015). Neighborhood deprivation and physical activity facilities: No support for the deprivation amplification hypothesis. Journal of Physical Activity and Health, 12(8), 1050–1057. https://www.researchgate.net/publication/265473694

[78] Office for National Statistics. (2022). Population estimates for the UK, England, Wales, Scotland and Northern Ireland: mid-2022. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2022

[79] Office for National Statistics. (2023). 2021 residential-based area classifications. https://www.ons.gov.uk/methodology/geography/geographicalproducts/areaclassifications/2021residentialbasedareaclassifications

[80] OpenStreetMap contributors. (n.d.). OpenStreetMap data for England [Data set]. Retrieved Mar 25, 2025, from Geofabrik: https://download.geofabrik.de/

[81] Ordnance Survey. (2022). Built-Up Areas (BUA), UK. https://osdatahub.os.uk/downloads/open/BUA

[82] Ordnance Survey. (2023a). OS Open Roads [Data set]. https://osdatahub.os.uk/downloads/open/OpenRoads

[83] Ordnance Survey. (2023b). OS Zoomstack [Data set]. https://osdatahub.os.uk/downloads/open/Zoomstack

[84] Perdue, W. C., Stone, L. A., & Gostin, L. O. (2006). The built environment and its relationship to the public's health: The legal framework. American Journal of Public Health, 96(4), 586–589. https://doi.org/10.2105/AJPH.2004.063321

[85] Peters, K., Elands, B., & Buijs, A. (2010). Social interactions in urban parks: Stimulating social cohesion? Urban Forestry & Urban Greening, 9(2), 93–100. https://doi.org/10.1016/j.ufug.2009.11.003

[86] Pohlmann, C., & Opitz, C. (2013). Typology of brand positioning strategies. Journal of Brand Management, 20(6), 465–476. https://doi.org/10.1057/bm.2012.50

[87] Project OSRM contributors. (n.d.). Open Source Routing Machine (OSRM). https://project-osrm.org/

[88] Public Accounts Committee. (2023). Grassroots participation in sport and physical activity. House of Commons. https://publications.parliament.uk/pa/cm5803/cmselect/cmpubacc/46/report.html

[89] Reed, J., & Buck, S. (2009). The effect of regular aerobic exercise on positive-activated affect: A meta-analysis. Psychology of Sport and Exercise, 10(6), 581–594. https://doi.org/10.1016/j.psychsport.2009.05.009

[90] Reyment, R. A., & Jöreskog, K. G. (1993). Applied factor analysis in the natural sciences (2nd ed.). Cambridge University Press. SpringerLink

[91] Robinson, E., Gillespie, S., & Jones, A. (2020). Obesity, eating behavior and physical activity during COVID-19 lockdown: A study of UK adults. Appetite, 156, 104853. https://doi.org/10.1016/j.appet.2020.104853

[92] Romp, W., & de Haan, J. (2007). Public capital and economic growth: A critical survey. Perspektiven der Wirtschaftspolitik, 8(S1), 6–52. https://doi.org/10.1111/j.1468-2516.2007.00242.x

[93] Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

[94] Sargent, T. J. (1993). Rational expectations and inflation (2nd ed.).

[95] Sport England, (n.d.-a) Active Places [Data set]. Retrieved Mar 25, 2025, from https://www.activeplacespower.com/

[96] Sport England. (2021). The impact of coronavirus on activity levels revealed. https://www.sportengland.org/news/impact-coronavirus-activity-levels-revealed

[97] Sport England. (2021). Inclusive physical activity and reducing inequalities: What we've learned from evaluation. https://sportengland-production-files.s3.eu-west-2.amazonaws.com/s3fs-public/2021-11/Inclusive%20physical%20activity%20and%20reducing%20inequalities%20-%20full%20report.pdf

[98] Sport England. (2022). The rising cost of living and its impact on sport and physical activity. https://sportengland-production-files.s3.eu-west-2.amazonaws.com/s3fs-public/2024-01/Cost%20of%20living%20impact%20report%20-%20full%20report%20%28January%202024%29.pdf

[99] Sport England. (n.d.-b). Facilities Planning Model. https://www.sportengland.org/how-we-can-help/facilities-and-planning/planning-for-sport/facilities-planning-model

[100] Sport England. (n.d.-c). Grants awarded [Data set]. 360Giving. https://grantnav.threesixtygiving.org/funder/GB-CHC-289548

[101] Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical Science, 25(1), 1–21. https://doi.org/10.1214/09-STS313

[102] Sutton, J. (1991). Sunk costs and market structure: Price competition, advertising, and the evolution of concentration. MIT Press.

[103] ter Braak, C. J. F. (1983). Principal components biplots and alpha and beta diversity. Ecology, 64(2), 454–462. https://doi.org/10.2307/1938291

[104] Train, K. E. (2009). Discrete choice methods with simulation (2nd ed.). Cambridge University Press.

[105] Tsou, Y. T., Lin, Y. C., & Lin, C. Y. (2022). Exploring the influence of social motivation on sports participation: A study of university students. Journal of Sports Science and Medicine, 21(1), 45–52.

[106] UCL News. (2024, September 4). 20% fall in GP surgeries while patient lists grow. https://www.ucl.ac.uk/news/2024/sep/20-fall-gp-surgeries-while-patient-lists-grow

[107] van den Dool, H. M. (2007). Empirical methods in short-term climate prediction. Oxford University Press.

[108] Varian, H. R. (1992). Microeconomic analysis (3rd ed.). W. W. Norton & Company.

[109] Viana, M. C., Gruber, M. J., Shahly, V., Alhamzawi, A. O., Alonso, J., Andrade, L. H., ... & Kessler, R. C. (2021). Family burden related to mental and physical disorders in the world: Results from the

WHO World Mental Health (WMH) surveys. Revista Brasileira de Psiquiatria, 43(1), 18–29. https://doi.org/10.1590/1516-4446-2020-1040

[110] Vu, V. Q., Cho, J., Lei, J., & Rohe, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In Advances in Neural Information Processing Systems, 26. https://proceedings.neurips.cc/paper_files/paper/2013/file/4b03bdfd3d9b8b1b0e1e5d2c1b1e5d2c-Paper.pdf

[111] Warburton, D. E. R., Nicol, C. W., & Bredin, S. S. D. (2006). Health benefits of physical activity: The evidence. Canadian Medical Association Journal, 174(6), 801–809. https://doi.org/10.1503/cmaj.051351

[112] Witten, D. M., Tibshirani, R., & Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics, 10(3), 515–534. https://doi.org/10.1093/biostatistics/kxp008

[113] Witten, D., Tibshirani, R., Gross, S., & Narasimhan, B. (2024). PMA: Penalized multivariate analysis (R package version 1.2-4). https://CRAN.R-project.org/package=PMA

[114] Woods, M. (2005). Rural geography: Processes, responses and experiences in rural restructuring. SAGE Publications.

[115] Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data (2nd ed.). MIT Press.

[116] World Health Organization. (2017). Depression and other common mental disorders: Global health estimates. https://www.who.int/publications/i/item/depression-global-health-estimates

[117] Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. Journal of Computational and Graphical Statistics, 15(2), 265–286. https://doi.org/10.1198/106186006X113430