Appendices

Appendix A

We selected the K-means algorithm for three considerations. First, the subsequent analytical framework requires each metro station to be assigned to a discrete land use pattern: we fit separate XGBoost models for each pattern (Section 4.2–4.3) to compare the relative importance and nonlinear effects of built environment variables across patterns. This stratified design is incompatible with the probabilistic membership vectors produced by soft clustering methods such as Gaussian Mixture Models (GMM). While GMM can represent transitional characteristics through partial membership, such vectors cannot be directly entered into pattern-specific models that assume discrete group membership. Converting probabilistic assignments to hard labels by selecting the maximum-probability cluster is feasible, but this discards GMM's core advantage while retaining its greater parametric complexity, yielding no analytical benefit over K-means for the downstream modeling task. Second, K-means imposes fewer parametric assumptions. GMM assumes within-cluster multivariate normality in the 12-dimensional feature space; with the smallest cluster containing 30 stations (transit hub pattern), reliable estimation of covariance parameters is infeasible (Hastie et al., 2009). K-means estimates only cluster centroids and makes no distributional assumptions about the within-cluster data-generating process, making it the methodologically conservative choice under modest sample sizes. Third, K-means is the standard method in transit-oriented development and land use typology studies (Liu et al., 2022; Zhang et al., 2019; Li et al., 2019; Kumar et al., 2020), ensuring comparability with established classification frameworks.

Appendix B

Conventional parametric models, including ordinary least squares regression, Poisson regression, geographically weighted regression (GWR), and multiscale geographically weighted regression (MGWR), have been widely applied in the M-DBS literature (Zhou et al., 2023; Yan et al., 2024; Gao, Hu, et al., 2023). These models, however, require pre-specified functional forms and thus cannot detect threshold effects or non-monotonic relationships (e.g., the inverted U-shaped association between shopping facility proportion and M-DBS usage observed in transit hub areas; Figure 11). Generalized Additive Models (GAMs; Wood, 2017) relax the linearity assumption by fitting smooth spline functions to each predictor and have been applied to capture nonlinear built environment effects in travel behavior research (Ding et al., 2019). Yet GAMs retain an additivity assumption: the effect of each variable is estimated independently of the levels of others. This is restrictive for built environment analysis, where interactions between variables, such as the dependence of population density effects on transit accessibility levels, are well documented (Gao, Yang, et al., 2023). Tree-based ensemble methods capture such interactions without requiring them to be pre-specified (Hastie et al., 2009) and are therefore better suited to our analytical objectives.

Among tree-based methods, the choice between bagging (Random Forest; Breiman, 2001) and boosting (XGBoost; Chen and Guestrin, 2016) carries distinct implications for small-sample settings. Random Forest constructs deep trees independently on bootstrap samples and averages predictions across the ensemble. This variance-reduction strategy is effective for large datasets, but two characteristics limit its suitability here. First, bootstrap sampling further reduces the effective training size of already small clusters, a concern most acute for the transit hub pattern ($n = 30$). Second, Random Forest controls model complexity only through tree depth and the number of trees, lacking an explicit mechanism to penalize overly complex individual trees (Hastie et al., 2009). XGBoost, by contrast, builds shallow trees sequentially, with each tree trained on the residual errors of the preceding ensemble. This sequential correction is more data-efficient than independent averaging because each iteration targets the observations that the current ensemble predicts most poorly (Friedman, 2001). Importantly, XGBoost incorporates L1 and L2 regularization terms on leaf weights (Equations 5–6), providing a formal

mechanism to constrain model complexity. In small-sample regimes, this regularization serves as the binding constraint against overfitting. The empirical results in Table 4 support this rationale: XGBoost achieves R^2 values of 0.83–0.86 across clusters, while Random Forest yields R^2 values of 0.60–0.71, with the largest performance gap observed in the smallest cluster (transit hub: 0.86 vs. 0.66).

Other gradient boosting implementations, such as LightGBM (Ke et al., 2017) and CatBoost (Prokhorenkova et al., 2018), have been proposed for specific computational scenarios. LightGBM employs leaf-wise tree growth and gradient-based one-side sampling, optimizations that accelerate training on large datasets but confer no advantage for our per-cluster sample sizes. Its leaf-wise growth strategy also tends to produce deeper trees, elevating overfitting risk when samples are limited (Ke et al., 2017). CatBoost's primary innovations ordered boosting to reduce target leakage and native categorical feature support, address problems less salient in our setting, where all 12 predictors are continuous (Prokhorenkova et al., 2018). XGBoost's depth-wise tree growth with explicit L1/L2 regularization provides the most conservative and well-regularized approach for small-sample regression, a setting in which controlling overfitting is the primary methodological concern.

References

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Ding, C., Cao, X., Dong, M., Zhang, Y., & Yang, J. (2019). Non-linear relationships between built environment characteristics and electric-bike ownership in Zhongshan, China. *Transportation Research Part D: Transport and Environment, 75*, 286–296. https://doi.org/10.1016/j.trd.2019.09.005

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gao, K., Yang, Y., Gil, J., & Qu, X. (2023). Data-driven interpretation on interactive and nonlinear effects of the correlated built environment on shared mobility. *Journal of Transport Geography, 110*, 103604. https://doi.org/10.1016/j.jtrangeo.2023.103604

Gao, W., Hu, X., & Wang, N. (2023). Exploring spatio-temporal pattern heterogeneity of dockless bike-sharing system: Links with cycling environment. *Transportation Research Part D: Transport and Environment, 117*, 103657. https://doi.org/10.1016/j.trd.2023.103657

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Spring Nature.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems 30*. https://proceedings.neurips.cc/paper_files/paper/2017

Kumar, P. P., Sekhar, C. R., & Parida, M. (2020). Identification of neighborhood typology for potential transit-oriented development. *Transportation Research Part D: Transport and Environment, 78*, 102186. https://doi.org/10.1016/j.trd.2019.11.015

Li, Z., Han, Z., Xin, J., Luo, X., Su, S., & Weng, M. (2019). Transit oriented development among metro station areas in Shanghai, China: Variations, typology, optimization and implications for land-use planning. *Land Use Policy, 82*, 269–282. https://doi.org/10.1016/j.landusepol.2018.12.003

Liu, S., Zhou, C., Rong, J., Bian, Y., & Wang, Y. (2022). Concordance between regional functions and mobility features using bike-sharing and land-use data near metro stations. *Sustainable Cities and Society, 84*, 104010. https://doi.org/10.1016/j.scs.2022.104010

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems 31*. https://proceedings.neurips.cc/paper_files/paper/2018

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279

Yan, Y., & Chen, Q. (2024). Spatial heterogeneity and nonlinearity study of bike-sharing to subway connections from the perspective of built environment. *Sustainable Cities and Society, 114*, 105766. https://doi.org/10.1016/j.scs.2024.105766

Zhang, Y., Marshall, S., & Manley, E. (2019). Network criticality and the node-place-design model: Classifying metro station areas in Greater London. *Journal of Transport Geography, 79*, 102485. https://doi.org/10.1016/j.jtrangeo.2019.102485

Zhou, X., Dong, Q., Huang, Z., Yin, G., Zhou, G., & Liu, Y. (2023). The spatially varying effects of built environment characteristics on the integrated usage of dockless bike-sharing and public transport. *Sustainable Cities and Society, 89*, 104348. https://doi.org/10.1016/j.scs.2022.104348