Smart Land Investment Assistant

Unified Product Document + Data/Modeling Framework

Version: v1.2 (Deduplicated, logically reordered) | Document language: English | Style: Startup/Product language, implementation-aligned

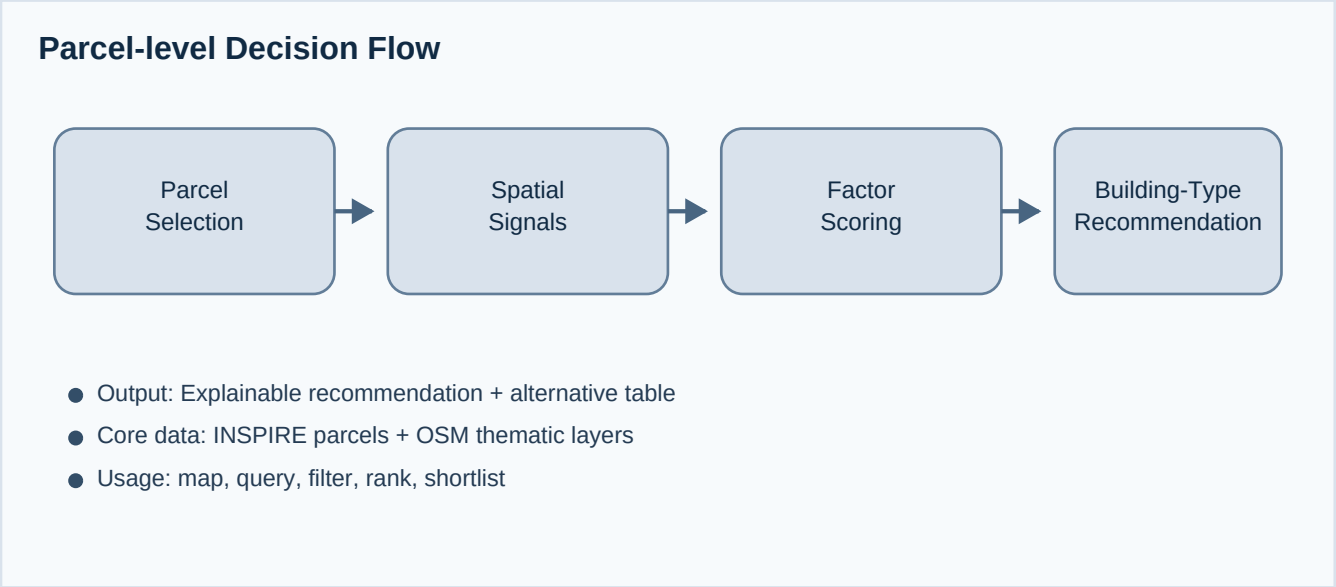

Figure 1. High-level parcel-to-recommendation decision flow.

Purpose: Provide parcel-level, explainable recommendation on which building type is more suitable and more profitable.

1. Product Summary

Smart Land is a parcel-level decision-support product that enables an investor to select a parcel on a map and get an explainable answer to: "Which building type is more suitable and more profitable on this parcel?"

The product combines parcel geometry with contextual spatial signals and produces a single analyzable output for mapping, querying, filtering, ranking, and shortlisting.

Core spatial integration is based on INSPIRE parcel geometry + OSM-derived thematic layers, with investment-oriented scoring and scenario comparison in the product layer [10, 6, 7, 14, 15, 16, 28, 29, 30].

2. Problem Definition

- Data fragmentation: planning, land, transport, and environmental datasets are spread across multiple sources [32].
- Decision bridge gap: raw spatial data does not directly answer "What should be developed here?"
- Manual GIS overhead: CRS conversion, geometry repair, clipping, and join steps are repeatedly rebuilt per project [34, 35, 36, 37].
- Inconsistent methods: team-level workflow differences reduce comparability.
- Performance risk: repeated live computations and overly granular category taxonomies increase latency [42, 43, 44].

3. Target Users and Jobs-to-be-Done

3.1 Primary users

- Individual investors: fast parcel-level building-type decisions
- Developers/contractors: return-oriented ranking across parcels/scenarios
- Consultants/valuation specialists: transparent, auditable reporting

- GIS/spatial analysts: integration, QA, repeatable geoprocessing
- Planning/operations teams: regional accessibility and capacity assessment

3.2 Jobs-to-be-done

- "Quickly identify high-potential parcels."
- "Compare parcels by proximity to schools, hospitals, safety, and transport."
- "Run a repeatable pipeline with consistent outputs across regions."

4. Value Proposition

- Fast decisions: recommendation + rationale in one flow
- Transparent scoring: explicit factor breakdowns
- Comparable analysis: same criteria across parcels and scenarios
- Operational repeatability: standardized CRS/schema and reproducible processing
- Business-ready outputs: map/table/export packages for GIS and non-GIS teams

5. Product Scope

5.1 In-scope (MVP + near-term)

A) Decision-support layer

- Parcel selection (map + parcel ID)
- Nearest units and distance table
- Alternative building-type table (full list)
- Factor-coefficient matrix scoring
- Sale-candidate and development-candidate tags
- Cross-parcel ranking by total investment score
- Basic scenario comparison (profitability + risk/return metrics)

B) Spatial pipeline layer (QGIS-first)

- INSPIRE + OSM integration [10, 6, 7]
- Shared analysis CRS: EPSG:27700 for meter-based distance analysis [24]
- Geometry QA/repair (Fix Geometries family) [36]
- AOI clipping (Clip / Extract by extent) [35]
- Parcel feature engineering: area/perimeter; nearest POI distance; POI counts in buffers; transport proximity metrics
- Enriched parcel output layer (GeoPackage/GeoJSON)

5.2 Out-of-scope (for now)

- Automated credit/financing recommendation
- End-to-end legal/title automation or title certainty verification
- Automated planning-permission generation
- Enterprise-grade full financial pro forma
- Automated AVM/ML value prediction model
- Full live web productization as primary delivery channel (MVP remains GIS-first for processing)

6. Success Metrics and KPIs

6.1 North Star metric

- Reduced time to produce analysis-ready parcel records (minutes per parcel)

6.2 Operational KPIs

- 50%+ reduction in data preparation time
- Repeatability: same input -> same output
- Geometry error rate: <1% after repair
- Missing mandatory feature rate: <5%

- Parcel click-to-insight response time (P95)

6.3 Business KPIs

- Time-to-shortlist reduced from days to hours
- Higher cross-team consistency via one data dictionary and one layer schema
- Increased usage of side-by-side parcel/scenario comparison views

6.4 Product usage/performance KPIs

- Parcel click -> panel response time (P95)
- Recommendation engine response time
- Parcels/scenarios compared per session
- Score table usage rate
- Click-through rate on recommended building type
- Weekly active scenario count

7. Data Strategy

7.1 Source classification (product UI language)

- Sale Candidate (Public Portfolio): Homes England Land Hub [14, 15]
- Development Candidate (Brownfield): Brownfield land / Brownfield site [16, 28, 29]
- Access & Amenities: OSM/Overpass + transport nodes + road/river layers [22, 23, 30, 26, 27]

Critical product rule:

- A Brownfield record does not mean guaranteed sale status or automatic market listing; this distinction must be explicit in the UI [16, 1].

7.2 Parcel-centric data model (minimum)

- parcel_id, geometry, area_m2, perimeter_m, centroid_lat, centroid_lon
- sale_signal, sale_status
- brownfield_signal, brownfield_part
- d_hospital, d_school, d_police, d_metro, d_bus, d_mall, d_pier, d_river, d_road
- factor_scores (F1...F10)
- building_type_scores: detached, apartment, residential complex, retail, office, mixed-use, industrial, agricultural
- For boundary-related layers, show an explicit "indicative extent" quality note [9, 31].

8. Spatial Data Processing and Matching Pipeline

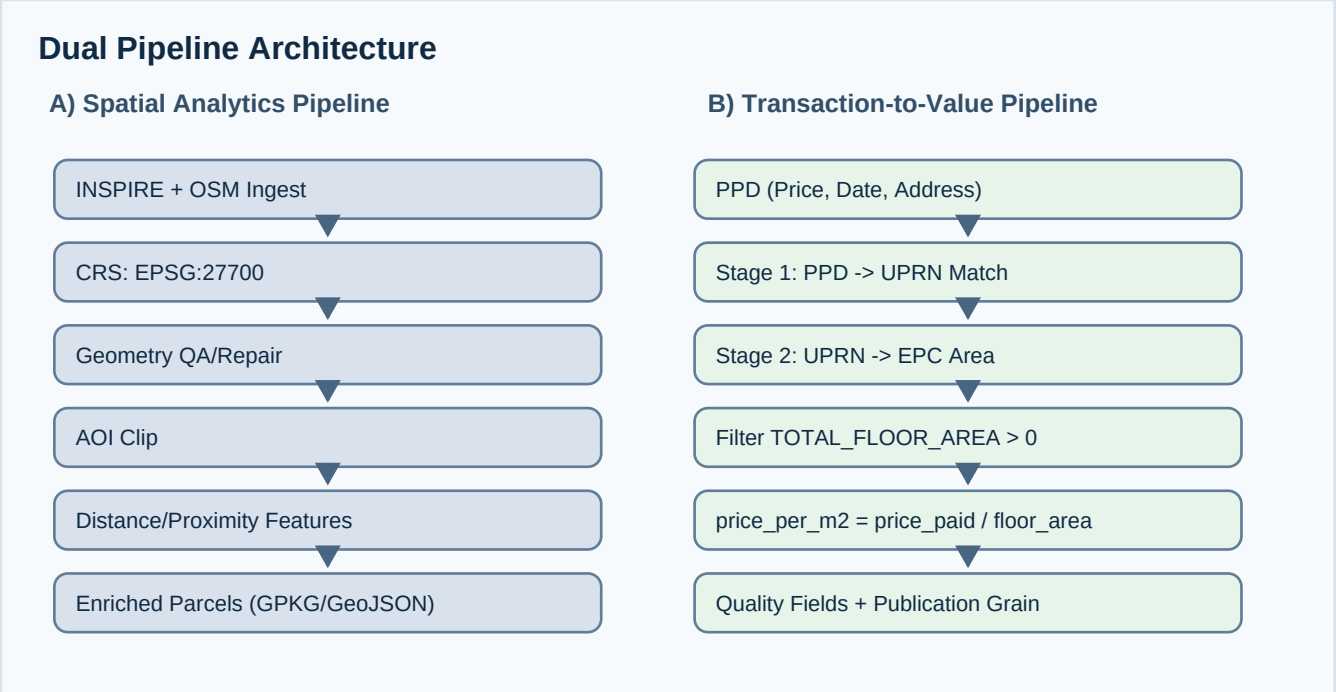

Figure 2. Two complementary pipelines: spatial feature engineering and transaction-normalized value modeling.

8.1 Spatial pipeline (QGIS-first)

- Ingest INSPIRE + OSM + supporting layers
- CRS standardization (analysis in EPSG:27700)
- Geometry validation/repair
- AOI clipping
- Distance and proximity feature engineering
- Enriched parcel output generation (GeoPackage/GeoJSON)
- This ensures repeatability and consistent metric semantics across regions.

8.2 Transaction/area pipeline for GBP/m2 modeling (Section 2 framework)

This second pipeline enriches parcel analytics with transaction-normalized value signals.

A) Dataset purposes

- Sale price core dataset - PPD: HM Land Registry Price Paid Data (PPD) is the primary transaction-price source with paid price, date, postcode, address parts, tenure, new-build flag, and transaction category; it covers value-based sales lodged for registration in England and Wales [11, 8, 13].
- m2 denominator + UPRN support - EPC: EPC provides TOTAL_FLOOR_AREA in square metres and includes UPRN; this is the core denominator for GBP/m2. EPC is not a full census, so outputs must include epc_coverage diagnostics [3, 4].
- Regional mapping - ONSPD: ONSPD maps postcode to geography (LA/LSOA/MSOA/region), standardizing aggregation and reducing boundary inconsistency [17].
- UPRN generation/support - OS Open UPRN: improves address matching robustness and geospatial join reliability [25].
- Commercial note - VOA: VOA rateable value is rental-value based, not sale price; therefore not directly comparable to sale-based GBP/m2. Commercial pipelines typically require additional proprietary inputs [45].

B) Two-stage UPRN matching strategy (production-grade)

- Because PPD has no direct UPRN field [8], use Stage 1 and Stage 2.
- Stage 1: PPD -> UPRN. Normalize postcode, PAON, SAON, street, town; deterministic match using exact postcode + PAON + SAON (+ street); fuzzy fallback within same postcode (threshold ~90-95 similarity); persist match_confidence (high/medium/low).
- Stage 2: UPRN -> EPC (m2). Handle multiple EPC certificates per UPRN; select EPC record nearest to sale_date by minimizing abs(sale_date - lodgement_date); filter invalid rows where TOTAL_FLOOR_AREA <= 0;

compute price_per_m2 = price_paid / total_floor_area.

• This preserves traceability, deterministic fallback behavior, and auditable quality labels [3, 8].

9. Decision Engine (Recommendation Engine)

Recommendation Engine: Factor x Building-Type Matrix

Factor	Detached	Apartment	Retail	Office	Mixed
F1	1	4	7	10	3
F2	3	6	9	2	5
F3	5	8	1	4	7
F4	7	10	3	6	9
F5	9	2	5	8	1
F6	1	4	7	10	3
F7	3	6	9	2	5
F8	5	8	1	4	7
F9	7	10	3	6	9
F10	9	2	5	8	1

Formula 1: U(p,t) = Sum_f [S(p,f) * W(t,f)] / Sum_f W(t,f)

Formula 2: I(p,t) = alpha*Sale + beta*Brownfield + gamma*U(p,t) (alpha=0.35, beta=0.20, gamma=0.45)

Figure 3. Example factor-by-building-type scoring matrix and final score formulas.

9.1 Factor set (F1-F10)

• F1 Zoning/planning

• F2 Legal constraints/encumbrances

• F3 Transport accessibility

• F4 Amenity proximity and walkability

• F5 Infrastructure/service capacity

• F6 Parcel geometry and size

• F7 Topography and ground conditions

• F8 Environmental constraints

• F9 Disaster hazard

• F10 Market and socioeconomics

9.2 Building-type coefficient matrix

• For each building type, coefficients from 0-10 are assigned across F1-F10, enabling comparability with a shared factor set.

9.3 Scoring formulas

• Parcel-factor score: S(p,f)

• Building-type coefficient: W(t,f)

• U(p,t) = Sum_f [S(p,f) * W(t,f)] / Sum_f W(t,f)

• Final investment score: I(p,t) = alpha*Sale + beta*Brownfield + gamma*U(p,t)

• Suggested initial weights: alpha=0.35, beta=0.20, gamma=0.45

10. GBP/m2 Publication Framework (Regional + Type)

10.1 Recommended analytical grain

• region / district / LSOA + property_type + year-quarter

10.2 Required output metrics

• Median GBP/m2

- IQR
- Observation count (n)

10.3 Outlier treatment

- Bottom/top 1% winsorization or IQR-fence filtering

10.4 Mandatory quality fields

- match_confidence
- days_diff_epc_sale
- epc_coverage_flag
- This is robust for residential segments. Commercial segments require additional explicit coverage diagnostics due to fragmented area/use/price signals [4, 45].

11. Modeling Framework for "Factor x Property-Type Coefficients"

11.1 Model 1 (interpretable): Hedonic with interactions

- Target: y = log(price_per_m2)
- Specification: y = beta0 + beta_type + Sum_f beta_f X_f + Sum_f beta_f,type (X_f x type) + u_region + u_t + epsilon
- Type-specific factor effect: coef_f,t = beta_f + beta_f,t
- This supports signed interpretation (positive/negative factor contribution by type).

11.2 Model 2 (predictive): LightGBM/CatBoost + SHAP

- Use same target log(price_per_m2), letting the model learn non-linear interactions.
- Type-conditional factor importance: S_f,t = mean(|SHAP_f| | type=t)
- Normalize to 1-10 scale: score_f,t = 1 + 9 * (S_f,t - min_t(S_f,t)) / (max_t(S_f,t) - min_t(S_f,t))
- This directly yields the required matrix format (factor rows x property-type columns).

12. Feature Families for Model Inputs

- Accessibility: distance to metro, distance to bus, arterial-road access
- Public services: school score/distance, hospital distance
- Market: neighborhood 12-month median GBP/m2, transaction volume
- Structural: building age, new-build flag, tenure, EPC band
- Legal/planning: conservation-area signal, planning-intensity signal
- Risk: flood risk, noise, air quality
- These are spatially engineered into a single analytical table before fitting.

13. Single-Property Quick Valuation ("Nearest m2 value")

13.1 Comparable-set constraints

- Same/similar property type
- Same tenure
- 500 m - 1 km radius
- +/-12/18-month transaction window

13.2 Weighting function

- w_i = exp(-d_i/lambda_d) * exp(-|delta_t_i|/lambda_t)

13.3 Estimate

- ppsm_hat = weighted_median(ppsm_i, w_i)
- This improves one-off valuation stability when comparables are sparse/noisy.

14. UI/UX Design

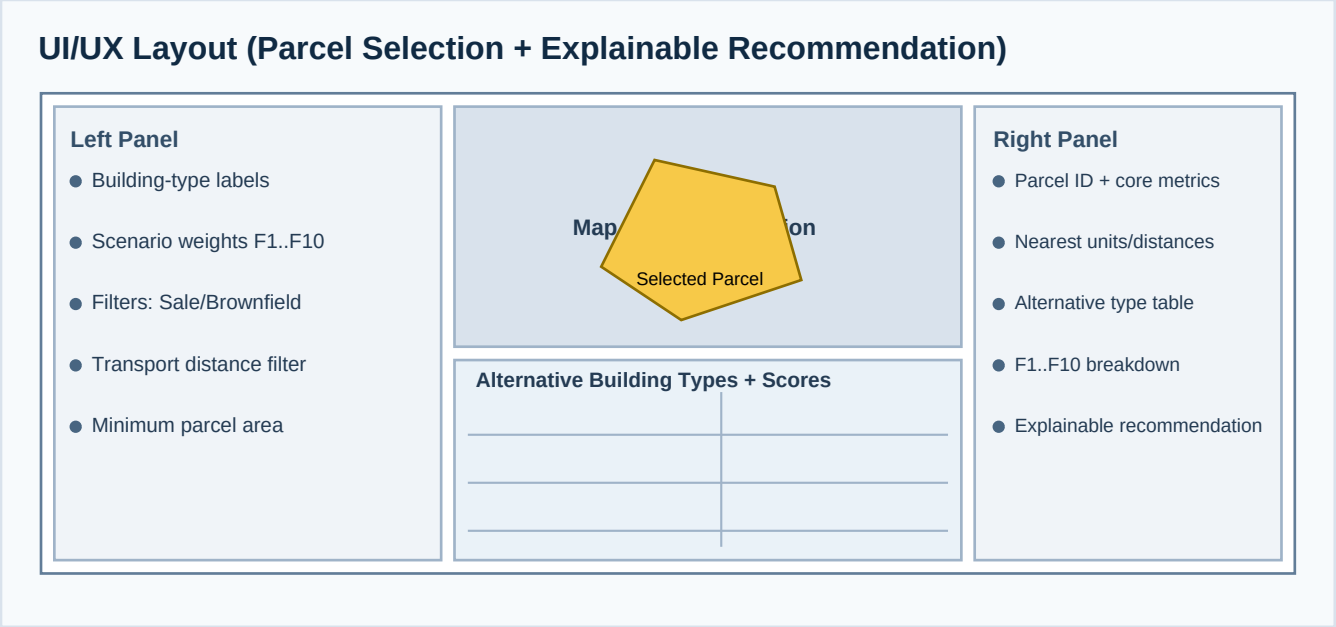

Figure 4. UI wireframe: left controls, map center, right explainability panel.

14.1 Left panel

- Building type labels
- Scenario weights (F1...F10)
- Filters: Sale candidates only; Brownfield only; Under X meters to public transport; Minimum parcel area

14.2 Right panel (selected parcel)

- Parcel ID and core metrics
- Nearest units and distance values
- Alternative building types table
- Factor breakdown (F1...F10)
- Explainable recommendation text (e.g., "Apartment score increased due to F3/F4/F10.")

15. Performance Architecture

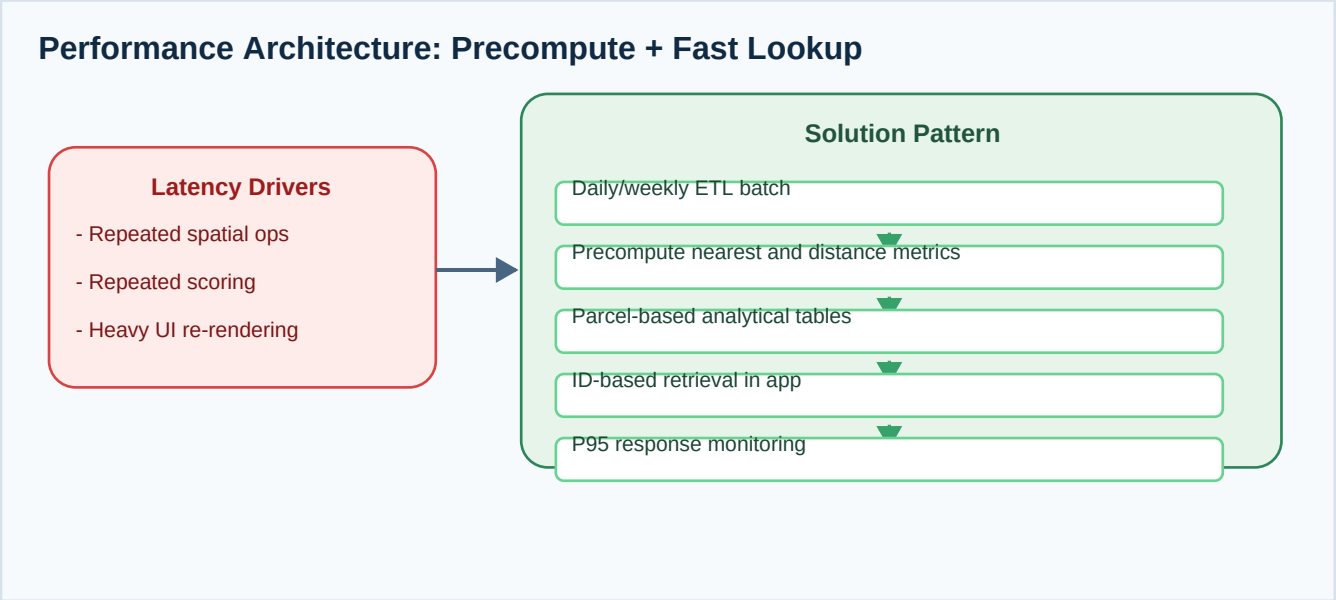

Figure 5. Operational pattern for low-latency parcel analytics.

15.1 Performance problem

- Repeated spatial operations, scoring calculations, and rendering per interaction create latency [42, 43, 44].

15.2 Solution principle

- Precompute + fast lookup instead of fully live computation:
- Daily/weekly ETL batch
- Precompute nearest/distance metrics
- Produce parcel-based analytical tables
- Use ID-based fast retrieval in app

15.3 Technical note

- Use GIS-native methods for distance/proximity operations [33, 38].

16. Trust and Accuracy Layer

- Must be visible in product outputs: source name; update date; quality label/flag; "Not for sale" warning for Brownfield; indicative boundary warning; "Stale" badge for old data.
- Because coverage/quality vary by source, a quality flag is mandatory [32].

17. Sprint Plan

Sprint 1

- Full building-type table
- F1...F10 breakdown
- Parcel-level score rows
- Explainable recommendation text

Sprint 2

- Offline precompute pipeline
- ready_parcel_metrics outputs
- Lookup-based speed optimization
- Performance logging

Sprint 3

- Sale/development badges
- Portfolio comparison screen
- Save/re-run scenario

18. Acceptance Criteria

- When a parcel is selected: nearest units + distance table must be visible.
- Full alternative building-types list must be visible.
- F1...F10 breakdown must be visible.
- Total investment score must be visible.
- P95 response target must be met in a 2,000-parcel scenario.
- Source label + Brownfield warning must be visible.

19. Technology Roadmap

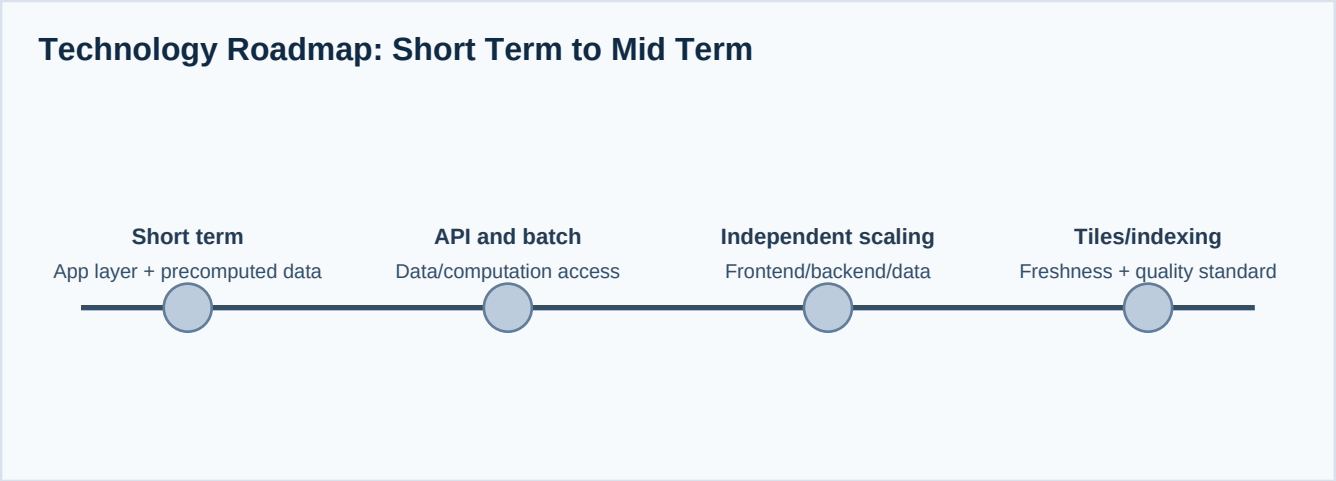

Figure 6. Short-term and mid-term technology roadmap milestones.

19.1 Short term

- Application layer + precomputed data layer
- API layer for data/computation access
- Batch update flow

19.2 Mid term

- Independent scaling of frontend/backend/data layers
- Geospatial tile/indexing approach
- Standardization of source freshness and quality labels

20. Critical Realism: Residential vs Commercial

- Residential classes are generally more feasible in open-data pipelines due to cleaner sale-price and labeling availability [11].
- For commercial classes, "sale price + floor area + clean use-type label" is fragmented in open data.
- Non-domestic EPC may provide area, while VOA provides rateable value (not sale price).
- Therefore, commercial sale-GBP/m2 modeling typically requires premium/proprietary enrichment [3, 45].

21. Implementation Alignment Check (Current MVP)

21.1 Implemented in current app

- Distance-based amenity scoring and nearest-unit logic
- Factor-by-building-type table editing in UI
- Scenario generation and financial outputs (profit, ROI, annualized IRR)
- Precompute/caching-oriented performance paths (prepared/simplified Parquet usage)

21.2 Partially implemented

- Factor x property-type matrix exists in UI/heuristic form, but not yet calibrated from an externally transaction-trained model
- Spatial-data workflow exists, but not yet wired to the full national open-data ingestion chain described above

21.3 Not yet implemented from this framework

- Direct ingestion pipeline for PPD + EPC + ONSPD + OS Open UPRN
- Two-stage production UPRN matching with persisted confidence labels
- Region/property-type/year-quarter GBP/m2 publication layer with full quality diagnostics
- Hedonic interaction model and LightGBM/CatBoost+SHAP training in production flow
- Commercial-grade enrichment beyond open-data baseline
- Conclusion of alignment: The MVP is architecturally compatible with this methodology; however, the full data-engineering and model-training stack is a next-phase production roadmap, not yet fully operational.

22. Positioning Statement

Smart Land is a map-based decision-support product that shows investors not only where a parcel is located, but also which building type can convert it into higher value in a numeric and explainable way.

References

[1] Bedford Borough Council. (n.d.). Brownfield land register. Bedford Borough Council. Retrieved February 9, 2026, from https://www.bedford.gov.uk/planning-and-building-control/planning-policy/brownfield-land-register

[2] Butler, H., Daly, M., Doyle, A., Gillies, S., Hagen, S., & Schaub, T. (2016). The GeoJSON format (RFC 7946). Internet Engineering Task Force. https://www.rfc-editor.org/rfc/rfc7946

[3] Department for Levelling Up, Housing and Communities. (n.d.-a). CSVW table schemas. Energy Performance of Buildings Open Data. Retrieved February 9, 2026, from https://epc.opendatacommunities.org/docs/csvw

[4] Department for Levelling Up, Housing and Communities. (n.d.-b). Energy Performance of Buildings documentation: Guidance. Energy Performance of Buildings Open Data. Retrieved February 9, 2026, from https://epc.opendatacommunities.org/docs/guidance

[5] EPSG Registry. (2021). OSGB36 / British National Grid (EPSG:27700). International Association of Oil & Gas Producers. https://epsg.org/crs_27700/OSGB36-British-National-Grid.html

[6] Geofabrik GmbH. (n.d.). Downloads [Service information].

[7] Geofabrik GmbH. (n.d.). England: OpenStreetMap data extracts. Retrieved February 9, 2026, from https://download.geofabrik.de/europe/united-kingdom/england.html

[8] HM Land Registry. (2016, June 14). How to access HM Land Registry Price Paid Data. GOV.UK. https://www.gov.uk/guidance/about-the-price-paid-data

[9] HM Land Registry. (2020, July 3). INSPIRE index polygons spatial data [Guidance]. GOV.UK. https://www.gov.uk/guidance/inspire-index-polygons-spatial-data

[10] HM Land Registry. (2021, September 16). INSPIRE Index Polygons spatial data [Dataset]. data.gov.uk.

[11] HM Land Registry. (2026, January 29). Price Paid Data. GOV.UK. https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads

[12] HM Land Registry. (n.d.). Download Index polygons spatial data (INSPIRE) [Data service]. Use Land and Property Data Service. https://use-land-property-data.service.gov.uk/datasets/inspire/download

[13] HM Land Registry Open Data. (n.d.). Search the price paid dataset. Retrieved February 9, 2026, from https://landregistry.data.gov.uk/app/ppd/

[14] Homes England. (2023, April 4). Using the Homes England Land Hub. GOV.UK. https://www.gov.uk/guidance/using-the-homes-england-land-hub

[15] Homes England. (n.d.). Land Hub sites [Dataset]. data.gov.uk. Retrieved February 9, 2026, from https://www.data.gov.uk/dataset/630b2ebc-157d-44e2-9a6e-2cf9cc8ea424/land-hub-sites

[16] Ministry of Housing, Communities and Local Government. (2017, July 28). Brownfield land registers. GOV.UK. https://www.gov.uk/guidance/brownfield-land-registers

[17] Office for National Statistics. (2025). ONS Postcode Directory (November 2025) for the UK (Hosted Table). data.gov.uk. https://www.data.gov.uk/dataset/56a898b9-8a3f-4c5f-8932-a03d958e0a43/ons-postcode-directory-november-2025-for-the-uk-hosted-table

[18] Office for National Statistics. (2025). Local Authority Districts (May 2025) Boundaries UK BFE (V2) [Dataset]. data.gov.uk.

[19] Open Geospatial Consortium. (2023). OGC GeoPackage encoding standard (Version 1.4.0). https://www.ogc.org/standard/geopackage/

[20] Open Geospatial Consortium. (n.d.). GeoPackage. Retrieved February 9, 2026, from https://www.ogc.org/standards/geopackage/

[21] OpenStreetMap Foundation. (n.d.). Copyright and license. Retrieved February 9, 2026, from https://www.openstreetmap.org/copyright

[22] OpenStreetMap Wiki contributors. (n.d.-a). Map features. OpenStreetMap Wiki. Retrieved February 9, 2026, from https://wiki.openstreetmap.org/wiki/Map_features

[23] OpenStreetMap Wiki contributors. (n.d.-b). Overpass turbo/GeoJSON. OpenStreetMap Wiki. Retrieved February 9, 2026, from https://wiki.openstreetmap.org/wiki/Overpass_turbo/GeoJSON

[24] Ordnance Survey. (2024, May 14). Coordinate reference systems. OS National Geographic Database documentation.

[25] Ordnance Survey. (2025). OS Open UPRN. data.gov.uk. https://www.data.gov.uk/dataset/c4f80d19-8cfa-4bf6-a283-83183842f876/os-open-uprn

[26] Ordnance Survey. (n.d.-a). OS Open Roads. OS Data Hub. Retrieved February 9, 2026, from https://osdatahub.os.uk/downloads/open/OpenRoads

[27] Ordnance Survey. (n.d.-b). OS Open Rivers. OS Data Hub. Retrieved February 9, 2026, from https://osdatahub.os.uk/downloads/open/OpenRivers

[28] Planning Data. (n.d.-a). Brownfield land [Dataset]. Retrieved February 9, 2026, from https://www.planning.data.gov.uk/dataset/brownfield-land

[29] Planning Data. (n.d.-b). Brownfield site [Dataset]. Retrieved February 9, 2026, from https://www.planning.data.gov.uk/dataset/brownfield-site

[30] Planning Data. (n.d.-c). Public transport access node [Dataset]. Retrieved February 9, 2026, from https://www.planning.data.gov.uk/dataset/transport-access-node

[31] Planning Data. (n.d.-d). Title boundary [Dataset]. Retrieved February 9, 2026, from https://www.planning.data.gov.uk/dataset/title-boundary

[32] Planning Data. (n.d.-e). Datasets. Retrieved February 9, 2026, from https://www.planning.data.gov.uk/dataset/

[33] PostGIS Project Steering Committee. (n.d.). ST_Distance. PostGIS Documentation. Retrieved February 9, 2026, from https://postgis.net/docs/ST_Distance.html

[34] QGIS Development Team. (n.d.-a). Working with projections. QGIS Documentation. https://docs.qgis.org/latest/en/docs/user_manual/working_with_projections/working_with_projections.html

[35] QGIS Development Team. (n.d.-b). Vector overlay: Clip / Extract by extent. QGIS Documentation. https://docs.qgis.org/latest/en/docs/user_manual/processing_algs/qgis/vectoroverlay.html

[36] QGIS Development Team. (n.d.-c). Vector geometry: Fix geometries. QGIS Documentation. https://docs.qgis.org/latest/en/docs/user_manual/processing_algs/qgis/vectorgeometry.html

[37] QGIS Development Team. (n.d.-d). Vector general: Join attributes by nearest / Join attributes by location. QGIS Documentation. https://docs.qgis.org/latest/en/docs/user_manual/processing_algs/qgis/vectorgeneral.html

[38] QGIS Development Team. (n.d.-e). Vector analysis: Distance matrix / Distance to nearest hub. QGIS Documentation. https://docs.qgis.org/latest/en/docs/user_manual/processing_algs/qgis/vectoranalysis.html

[39] QGIS Development Team. (n.d.-f). Vector general: Merge vector layers. QGIS Documentation. https://docs.qgis.org/latest/en/docs/user_manual/processing_algs/qgis/vectorgeneral.html

[40] QGIS.org. (n.d.-e). Vector geometry: Add geometry attributes. QGIS Documentation.

[41] QGIS.org. (n.d.-f). Working with joined tables and relations. QGIS Documentation.

[42] Streamlit. (n.d.-a). st.cache_data. Streamlit Docs. Retrieved February 9, 2026, from https://docs.streamlit.io/develop/api-reference/caching-and-state/st.cache_data

[43] Streamlit. (n.d.-b). Understanding widget behavior. Streamlit Docs. Retrieved February 9, 2026, from https://docs.streamlit.io/develop/concepts/architecture/widget-behavior

[44] Streamlit. (n.d.-c). Working with fragments. Streamlit Docs. Retrieved February 9, 2026, from https://docs.streamlit.io/develop/concepts/architecture/fragments

[45] Valuation Office Agency. (n.d.). How non-domestic property, including plant and machinery, is valued. GOV.UK. Retrieved February 9, 2026, from https://www.gov.uk/guidance/how-non-domestic-property-including-plant-and-machinery-is-valued