ORDNANCE SURVEY

Retail Geographies - High Streets Dataset Technical Specification

March 2019

Ordnance Survey

Copyright Notice

Version History

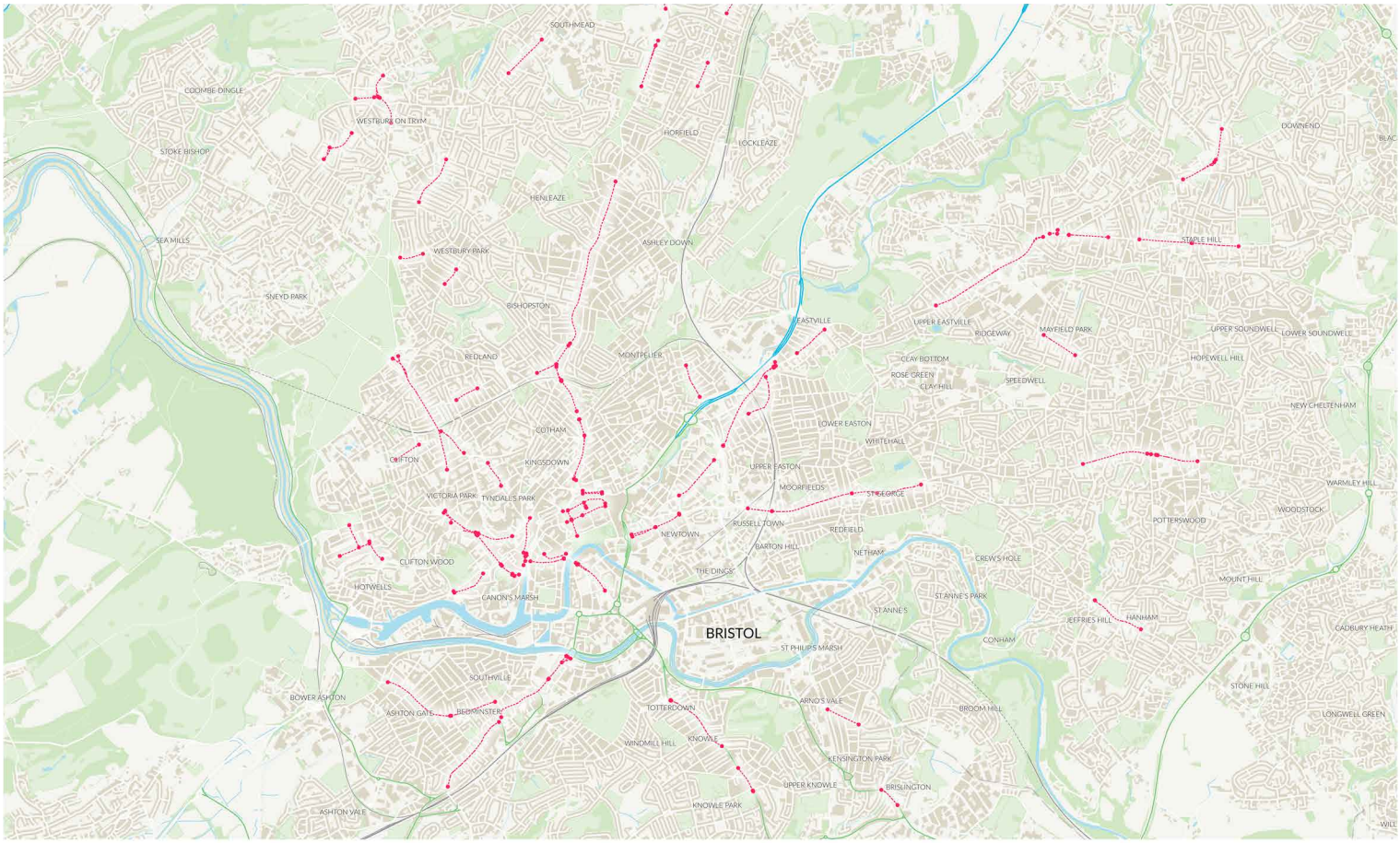
Version	Date	Description	Author
1.0	03/2019	High Streets Technical Specification	Steven Kingston, Data Scientist

Contents

1. Introduction

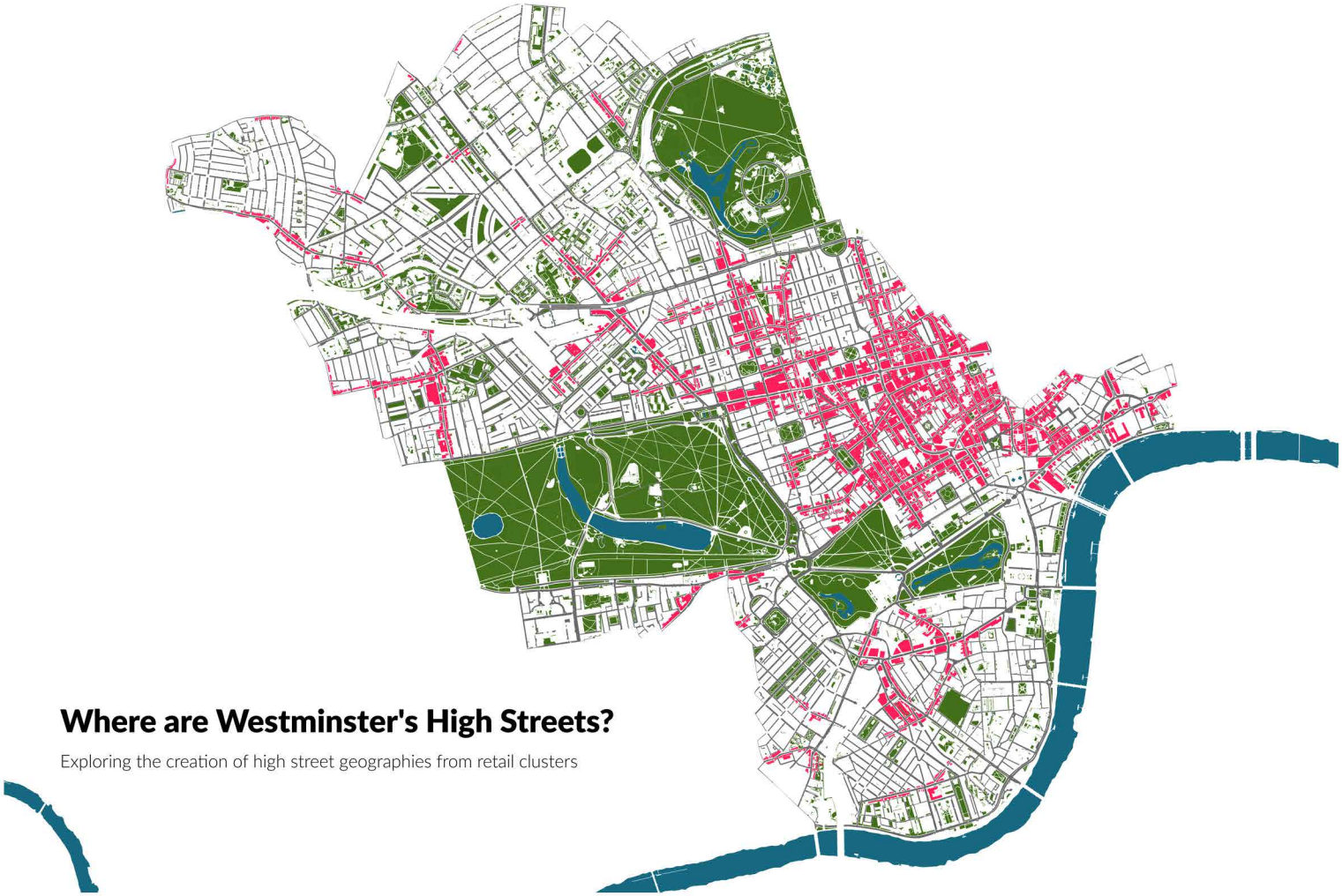
The March 2019 release of the high street dataset is the third iteration of an Ordnance Survey-derived dataset aiming to describe the spatial extent and profile of high streets across Great Britain. The March 2019 release builds on the previous two iterations improving the coverage and definition of high street geographies. The dataset is a high street-specific dataset that provides a street-level definition. It bounds retail clusters using street naming and deliberately excludes other retail functions including retail parks, industrial estates, and isolated shopping centres.

The dataset is derived from data products available under the Public Sector Mapping Agreement (PSMA) and One Scotland Mapping Agreement (OSMA), OS AddressBase Plus (Epoch 64), OS MasterMap Topography Layer (extraction date 2019-01-03 publication date 2019-01-21), OS MasterMap Sites Layer (October 2018) and OS VectorMap Local (April 2018 and January 2019). Due to the licensing terms of the source data used to derive the high street geography the dataset is only available to members of the PSMA or OSMA.



Where are Bristol's High Streets?

Exploring the creation of high street geographies from retail clusters



Where are Westminster's High Streets?

Exploring the creation of high street geographies from retail clusters

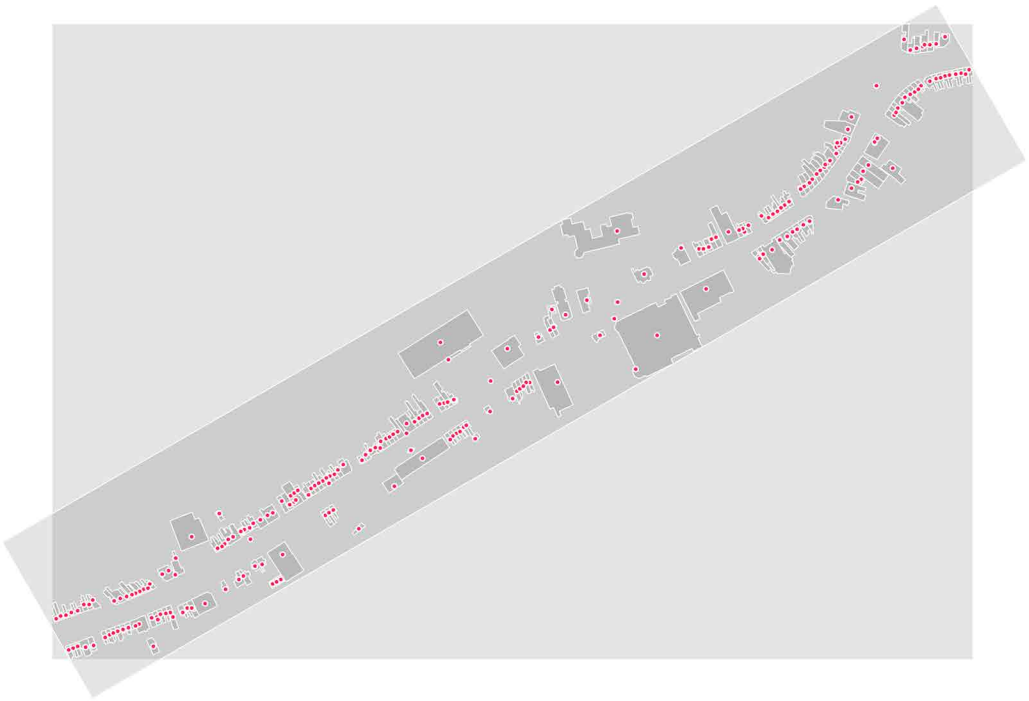
2. Data Coverage and Format

The high street data has been made available to download for all of Great Britain as well as by local authority district in a range of geospatial vector formats. For Great Britain-wide level, users can download the data in GeoPackage, GeoJSON, ESRI Shapefile and PostgreSQL formats. For local authority level, users can download the data in GeoPackage, GeoJSON and ESRI Shapefile formats.

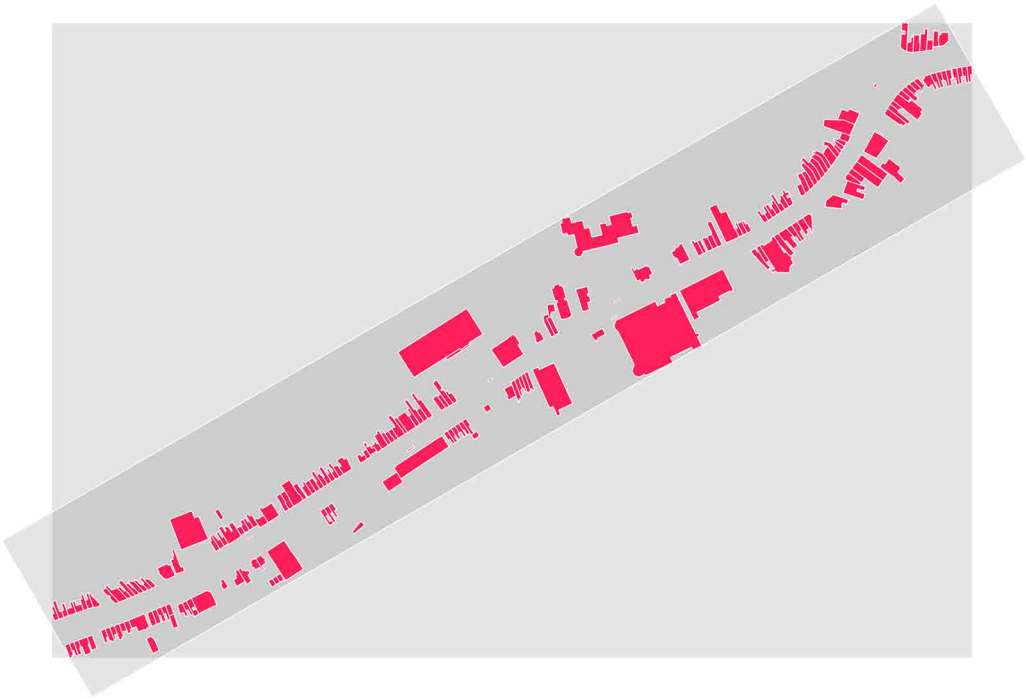
2.1 Geometry

The spatial extent of a high street is provided using four different geometries.

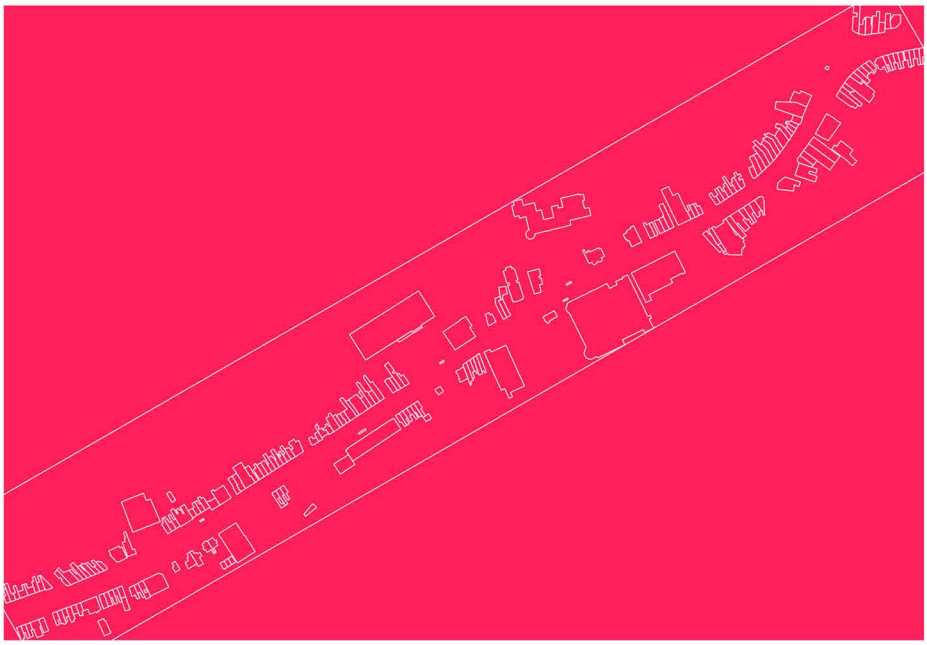
The address_geom provides a multipoint geometry collecting all of the address point geometries associated with a high street extent.



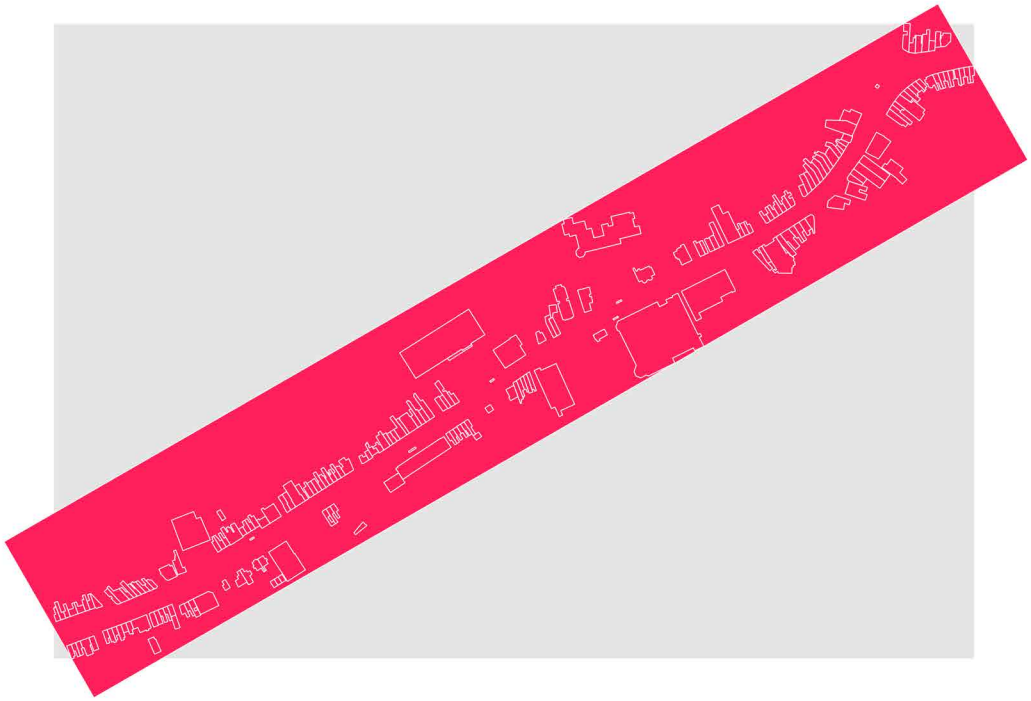
The building_geom provides a multipolygon geometry collecting all of the distinct building polygon geometries associated with a high street extent.



The bbox_geom provides an axis aligned bounding box or minimum bounding rectangle around the building_geom for a high street extent.



The orientated_bbox_geom provides a true, non-axis aligned bounding box or minimum bounding rectangle around the building_geom for a high street extent.



2.2 Lookups

The attributes associated with the address records which make up a high street extent have been used to derive a series of lookups which relate a high street geography to a number of statistical and geographical identifiers. The lookups are provided using array fields within the GeoJSON and PostgreSQL formats, as individual layers within the GeoPackage format, and independently of any of the spatial formats via comma separated value (CSV) text files.

highstreet_201903_id2uprn relates the value provided in the id field of a high street record with the Unique Property Identifier (UPRN) values for the addresses associated with that high street.

highstreet_201903_id2toid relates the value provided in the id field of a high street record with the distinct Topographic Identifier (TOID) values for the buildings associated with that high street.

highstreet_201903_id2usrn relates the value provided in the id field of a high street record with the distinct Unique Street Reference Number (USRN) values for the street records referenced by the addresses associated with that high street.

highstreet_201903_id2postcode relates the value provided in the id field of a high street record with the distinct postcode values for the addresses associated with that high street.

highstreet_201903_id2postcode_coordinates relates the value provided in the id field of a high street record with the distinct postcode values and their British National Grid easting northing coordinates for the addresses associated with that high street.

2.3 GeoPackage

The GeoPackage format option provides all four high street geometries and the four lookups as separate layers or tables within a single SQLite database container.

2.4 GeoJSON

The GeoJSON format option provides the four high street geometries as separate GeoJSON files. The lookups are provided using array fields within each GeoJSON file. Please see the attribute description section below for more information.

2.5 ESRI Shapefile

The ESRI Shapefile format option provides the four high street geometries as separate ESRI Shapefiles. The lookups are not provided as part of this option. Therefore, in order to relate the high street geographies to statistical and geographic identifiers the ESRI Shapefile option will need to be used in conjunction with the text file lookups.

2.6 PostgreSQL - PostGIS

The PostgreSQL format option provides a dump or backup of a database table containing the high street geography for Great Britain. The dump has been created in a custom dump format and when restored, will recreate the database table within the public database schema in the same state as it was at the time of the dump. The lookups are provided using array fields and all four geometries are included in the table using separate geometry fields. The PostGIS database extension is required to be created in the database which the restore targets.

The dump can be restored by running the pg_restore command which resides inside the bin directory where the PostgreSQL database server was installed.

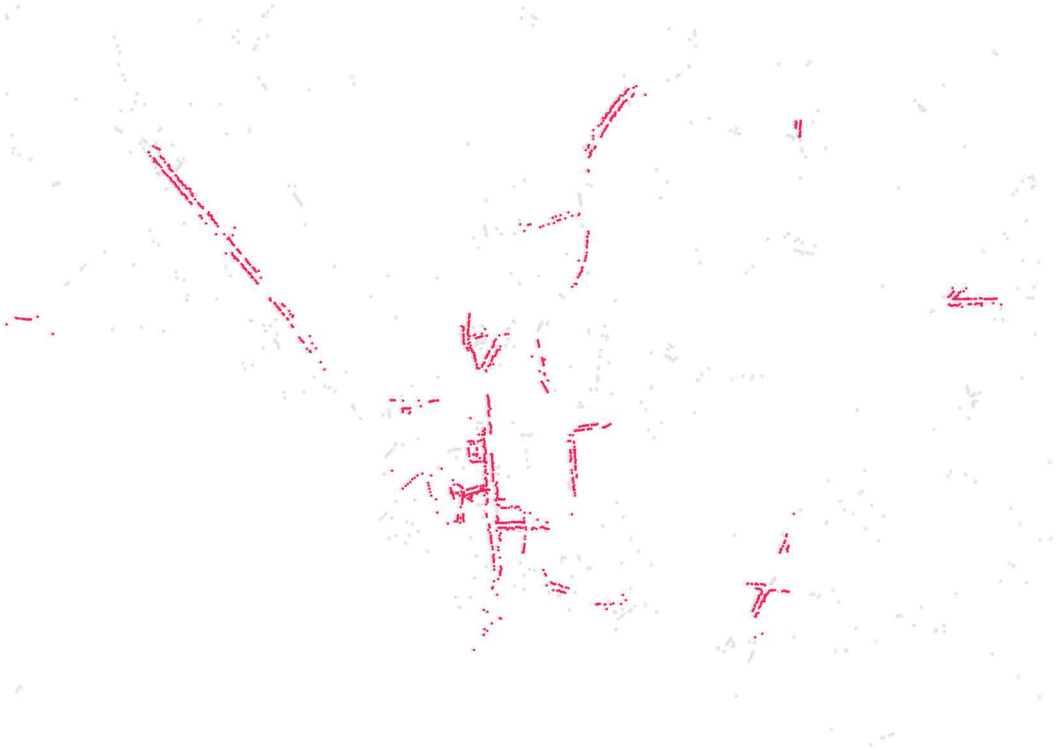
Use of pg_restore takes the following form:

pg_restore -j <number_of_jobs> -U <database_username> -h <host> -p <port> -d <database_name> <filename>

The dump file does not require to be uncompressed before it is restored.

3. Derivation

The high street dataset considers a high street as an extent along a named street predominately consisting of retailing. Spatial clustering of retail address point geometries from the product OS AddessBase Plus on a street-by-street basis was used to form candidate retail clusters from which a high street extent could be derived. Spatial clustering has been implemented using the data clustering algorithm DBSCAN (Density-based spatial clustering of applications with noise). Candidate retail clusters required a minimum of 15 retail address point geometries and a maximum of 150 metres between the point geometries in the cluster.



Insights relating to the building polygon geometries from the product OS MasterMap Topography Layer associated with the retail address geometries in each cluster were subsequently used to rule candidate clusters in or out of the dataset. For each cluster, the number of building geometries was evaluated as well as the length of the longest line passing through the cluster building multipolygon geometry. Clusters with less than 15 buildings, or a longest line length of less than 100 metres were removed from the dataset. The building count threshold of 15 was chosen so that it aligns with the minimum points parameter value passed to DBSCAN.

In order to evaluate the land use profile of each high street extent the retail clusters were expanded or infilled to collect all address point geometries, and their associated building polygon geometries, on the same named street within 50 metres of the building multipolygon geometry of a retail cluster. All high street attributes are derived from the infilled clusters and not the original retail clusters.

Exclusion of other non-high street retail functions

Other retail functions such as retail parks, industrial estates and isolated shopping centres have deliberately been excluded from the dataset using four tests. Each test evaluates each cluster against a combination of

form and function criteria. Whilst every effort has been made to exclude other retail functions we are aware that there are a small number of non-high street geographies remaining in the March 2019 release of the dataset.

Test 1: Text-based classification of properties in a cluster using their concatenated address strings.
This test generated a binary classification for each property address in a given cluster according to whether the concatenated address contained a banned substring. Banned substrings include combinations of words indicating association with non-high street retail functions such as retail parks, industrial estates, business parks and trading estates. Retail clusters where more than fifty percent of the property addresses contained a banned substring were removed from the dataset. The fourteen banned substrings are described in the table below. The list does not include the substring 'shopping centre' as high streets with shopping centres along their length were chosen not to be removed.

Banned Substring	
retail park	retail world
retail centre	shopping park
shopping village	industrial estate
industrial park	industrial centre
ind est	ind pk
trading estate	trading park
business centre	business park

Test 2: Text-based classification of the street name associated with a cluster.
This test generated a binary classification for a given cluster according to whether the street name contained a banned substring. Banned substrings include combinations of words indicating association with non-high street retail functions such as retail parks, industrial estates, business parks and trading estates. Retail clusters where the street name contained a banned substring were removed from the dataset. The banned substrings used in test 2 were the same as in test 1 apart from the substring 'shopping centre' which was introduced only for test 2. The substring 'shopping centre' was introduced as the occurrence of that substring in the street name is indicative of a dedicated shopping centre which can be isolated from other high street extents.

Test 3: Text-based classification of cartographic text seeds, cluster geometry compactness theshold and building-to-address count ratio threshold.
This test used three measures to identify non-high street retail clusters. The first measure identified retail clusters that were within 50 metres of cartographic text point geometry from the products OS VectorMap Local and OS MasterMap Topography Layer containing a banned substring. Banned substrings include combinations of words indicating association with non-high street retail functions such as retail parks, industrial estates, business parks and trading estates. The banned substrings used in test 3 were the same as in test 1. The second measure evaluated the compactness of a given cluster's orientated bounding box polygon geometry using the Polsby-Popper Test. Typically, the shape of retail parks, industrial parks and trading estates tend towards a square or circular shape with high compactness and the shapes of high street extents tend towards a more elliptical shape with low compactness. The third measure calculated the

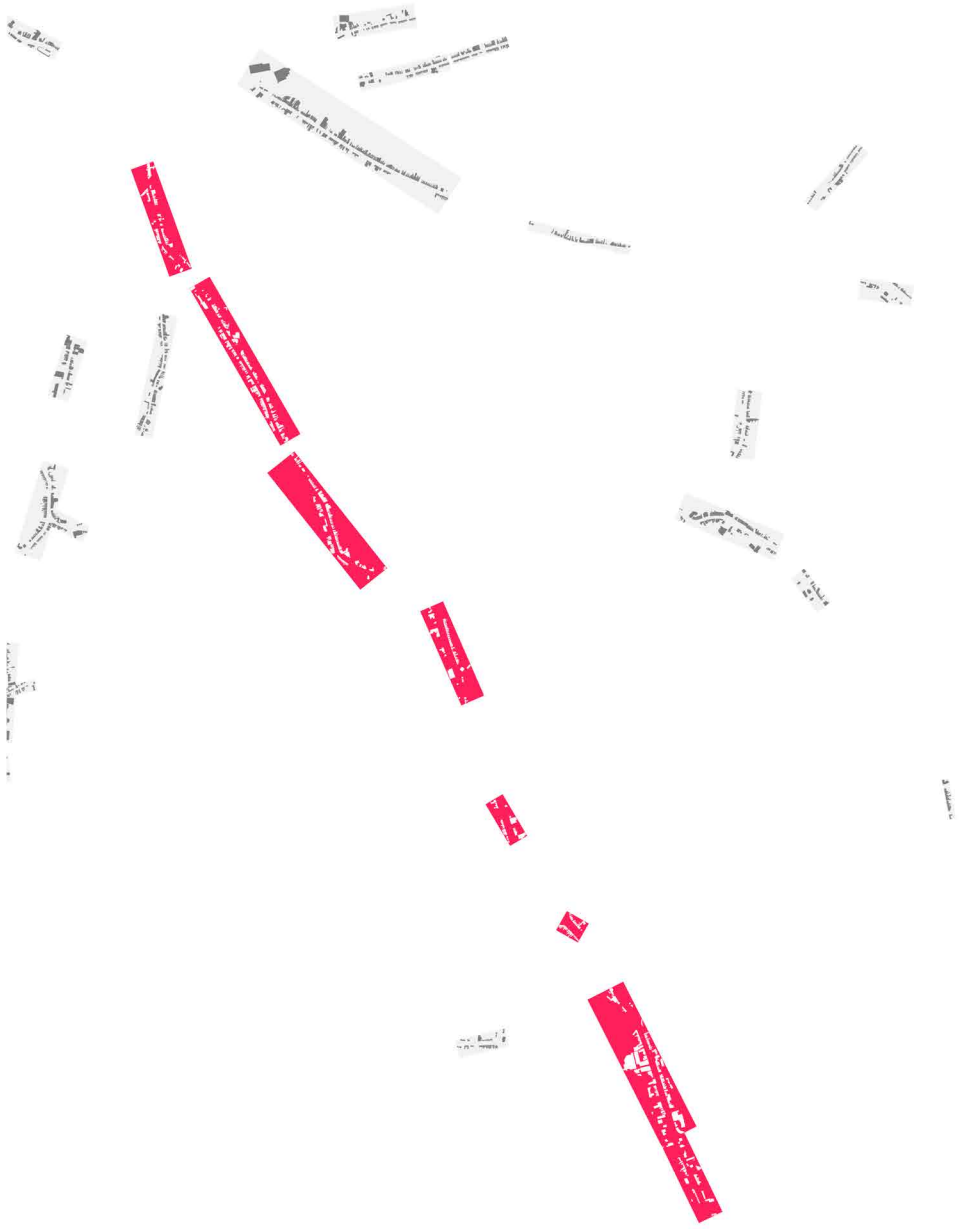
building-to-address ratio using a count of the building polygon geometries from OS MasterMap Topography Layer and a count of the address point geometries from OS AddressBase Plus. Typically, the buildings associated with retail parks, industrial parks and trading estates contain a single address or unit whereas the buildings associated with high street extents contain more than one address or unit. Retail clusters with a relationship to a banned cartographic text seed, a compactness score greater than or equal to 0.7 and a building-to-address count ratio of greater than or equal to 0.65 were removed from the dataset.

Test 4: Absence of residential land use, cluster geometry compactness threshold, building-to-address count ratio threshold and median building area threshold.
This test used four measures to identify non-high street retail clusters. The first measure identified retail clusters with a residential address count of zero. This measure was chosen since an absence of any residential land use is indicative of a non-high street retail function such as a retail park, industrial estate business park and trading estate. The second measure evaluated the compactness of a given cluster's orientated bounding box polygon geometry using the Polsby-Popper Test. Typically, the shape retail parks, industrial parks and trading estates tend towards a square or circular shape with high compactness and the shapes of high street extents tend towards a more elliptical shape with low compactness. The third measure calculated the building-to-address ratio using a count of the building polygon geometries from OS MasterMap Topography Layer and a count of the address point geometries from OS AddressBase Plus. Typically, the buildings associated with retail parks, industrial parks and trading estates contain a single address or unit whereas the buildings associated with high street extents contain more than one address or unit. The fourth measure evaluated the median building area using the building polygon geometries associated with a cluster from OS MasterMap Topography Layer. The median building area threshold was introduced since the typical building area of units in a retail park, industrial estate or trading estate is larger than the building area on most high street extents. Retail clusters with a no residential lands use, a compactness score greater than or equal to 0.7, a building-to-address count ratio of greater than or equal to 0.65 and a median building area greater than or equal to 250 square metres were removed from the dataset.

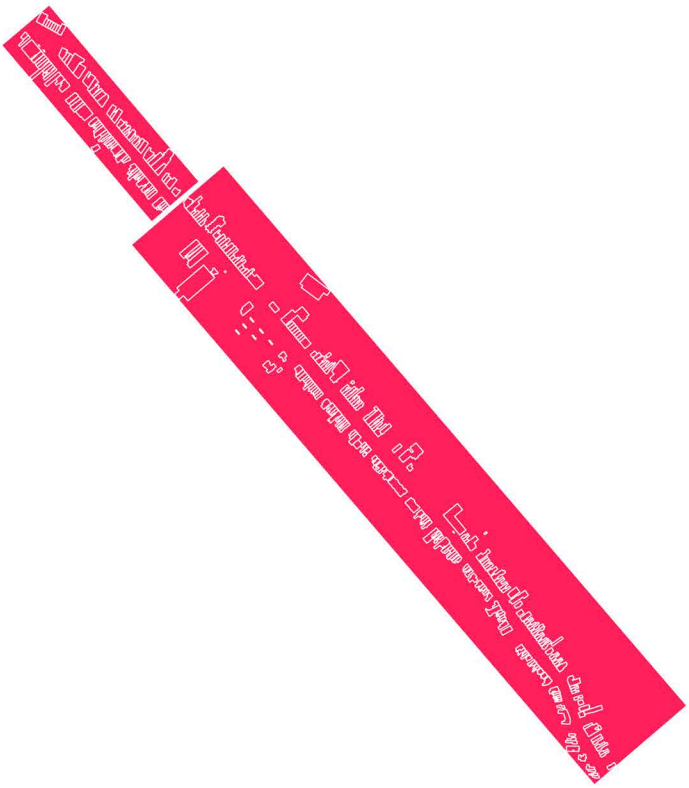
4. Data Features

This section presents insights into two notable data features that users should be aware of. Ordnance Survey would welcome feedback on these features in order to help assess whether future iterations of the high street dataset should handle them differently.

The first feature is where a named street has been split into more than one high street extent. This is because multiple dense retail clusters are spaced out along the length of the street. Below is an example in Birmingham, where Stratford Road has eight high street extents along its length.



The second feature is where seemingly one road is split into two different high street extents (even though the retail is continuous) because the addresses on one part of the road are assigned to a street name which is different to the other part. The example below is in Southampton, where the addresses on the north part of the road are assigned to Shirley High Street and the addresses on the south part to Shirley Road.



5. Attribute Descriptions

Field Name	Data Type	Availability	ESRI Shapefile Field Name
id	Character	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	id
Universally unique text-based identifier for each high street.			
street_description	Character	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	strt_desc
Name of a high street.			
street_length	Numeric	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	strt_lngth
Length of the longest line through the building multipolygon geometry that makes up a high street extent.			
building_count	Numeric	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	bdng_ct
Count of the polygon geometries classified as a building in the TopographicArea feature type of the product OS MasterMap Topography Layer that have been collected into the mutipolygon geometry associated with a high street extent.			
building_area_median	Numeric	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	bdng_a_med
Median building footprint area of the buildings comprising a high street extent to the nearest square metre.			
building_area_average	Numeric	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	bdng_a_avg
Average building footprint area of the buildings comprising a high street extent to the nearest square metre.			
building_area_max	Numeric	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	bdng_a_max
Maximum building footprint area of the buildings comprising a high street extent to the nearest square metre.			

Field Name	Data Type	Availability	ESRI Shapefile Field Name
address_count	Numeric	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	a_ct

Count of the address point geometries associated with a high street extent that have been mapped to one of the following Ministry of Housing, Communities and Local Government (MHCLG) land use categories *(for more information on the MHCLG land use categories please see Section 6. Notes)*:

- Retailing
- Offices
- Community Building
- Leisure and Recreational
- Residential

This count may be different to the number of address point geometries collected into the multipoint address geometry, and also different to the number of elements in the uprn array field, which references all of the Unique Property Reference Numbers (UPRNs) of the address records linked to a high street extent regardless of their classification.

retailing_address_count	Numeric	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	ret_a_ct

Count of the address point geometries where the classification value in the class field within the product OS AddressBase Plus has been mapped to the retailing MHCLG land use category.

offices_address_count	Numeric	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	of_a_ct

Count of the address point geometries where the classification value in the class field within the product OS AddressBase Plus has been mapped to the offices MHCLG land use category.

community_address_count	Numeric	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	comm_a_ct

Count of the address point geometries where the classification value in the class field within the product OS AddressBase Plus has been mapped to the community building MHCLG land use category.

leisure_and_recreational_address_count	Numeric	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	l_r_a_ct

Count of the address point geometries where the classification value in the class field within the product OS AddressBase Plus has been mapped to the leisure and recreational MHCLG land use category.

Field Name	Data Type	Availability	ESRI Shapefile Field Name
residential_address_count	Numeric	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	rs_a_ct
Count of the address point geometries where the classification value in the class field within the product OS AddressBase Plus has been mapped to the residential MHCLG land use category.			
retailing2offices_residential_address_count_ratio	Numeric	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	ret2of_res
Provides the retailing_address_count to the sum of offices_address_count and residential_address_count ratio.			
residential_house_address_count	Numeric	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	rs_h_a_ct
Identifies the subset of the residential_address_count which are classified as houses. Provides a count of the address point geometries mapped to the residential MHCLG land use category where the building polygon geometry which the address geometry spatially intersects does not have a relationship with any other address, residential or non-residential.			
residential_flat_address_count	Numeric	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	rs_f_a_ct
Identifies the subset of the residential_address_count which are classified as flats. Provides a count of the address point geometries mapped to the residential MHCLG land use category where the building polygon geometry which the address geometry spatially intersects does have a relationship with other addresses, residential or non-residential.			
residential_flat_mixed_address_count	Numeric	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	rs_fm_a_ct
Identifies the subset of the residential_flat_address_count which occupy a building with mixed land use, residential and non-residnential. Provides a count of the address point geometries mapped to the residential MHCLG land use category where the building polygon geometry which the address geometry spatially intersects does have a relationship with other non-residential addresses.			
building2address_count_ratio	Numeric	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	bdng2a_ct
Provides the building count to address_count ratio.			
lad_code	Array - Character	GeoJSON and PostgreSQL	

Field Name	Data Type	Availability	ESRI Shapefile Field Name
An array of the local authority district codes for the address custodians responsible for the maintenance of the address records associated with a high street extent. The elements in the array are provided in descending order based upon a count of the address records by local authority district. For a high street extent with a relationship to many local authority districts the ordering ensures that the most 'significant' district is provided first.			
lad_name	Array - Character	GeoJSON and PostgreSQL	
An array of the local authority district names ordered according to the elements within the lad_code array field, for the address custodians responsible for the maintenance of the address records associated with a high street extent. The elements in the array are provided in descending order based upon a count of the address records by local authority district. For a high street extent with a relationship to many local authority districts the ordering ensures that the most 'significant' district is provided first.			
lad_code1	Character	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	lad_code1
First element of the lad_code array. For a high street extent with a relationship to many local authority districts the lad_code1 field provides the code of the most 'significant' local authority district.			
lad_name1	Character	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	lad_name1
Provides the first element of the lad_name array. For a high street extent with a relationship to many local authority districts the lad_name1 field provides the name of the most 'significant' local authority district.			
lad_code2	Character	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	lad_code2
Second element of the lad_code array, if it exists.			
lad_name2	Character	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	lad_name2
Second element of the lad_name array, if it exists.			
lad_code3	Character	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	lad_code3
Third element of the lad_code array, if it exists.			
lad_name3	Character	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	lad_name3
Third element of the lad_name array, if it exists.			

Field Name	Data Type	Availability	ESRI Shapefile Field Name
region_code	Array - Character	GeoJSON and PostgreSQL	
An array of the region codes for the local authority districts of the address custodians responsible for the maintenance of the address records associated with a high street extent. The elements in the array are provided in descending order based upon a count of the address records by region. For a high street extent with a relationship to many regions the ordering ensures that the most 'significant' region is provided first.			
region_name	Array - Character	GeoJSON and PostgreSQL	
An array of the region names for the local authority districts of the address custodians responsible for the maintenance of the address records associated with a high street extent. The elements in the array are provided in descending order based upon a count of the address records by region. For a high street extent with a relationship to many regions the ordering ensures that the most 'significant' region is provided first.			
region_code1	Character	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	rgn_code1
First element of the region_code array. For a high street extent with a relationship to many regions the region_code1 field provides the code of the most 'significant' region.			
region_name1	Character	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	rgn_name1
First element of the region_name array. For a high street extent with a relationship to many regions the region_name1 field provides the name of the most 'significant' region.			
region_code2	Character	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	rgn_code2
Second element of the region_code array, if it exists.			
region_name2	Character	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	rgn_name2
Second element of the region_name array, if it exists.			
country_code	Character	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	ctry_code
Country code for the region(s) and local authority district(s) of the address custodians responsible for the maintenance of the address records associated with a high street extent.			
country_name	Character	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	ctry_name

Field Name	Data Type	Availability	ESRI Shapefile Field Name
Country name for the region(s) and local authority district(s) of the address custodians responsible for the maintenance of the address records associated with a high street extent.			
uprn	Array - Numeric	GeoJSON and PostgreSQL	
An array of the UPRNs identifying all of the address records associated with a high street extent. The number of elements in the uprn array is equal to the number of address point geometries collected into the address multipoint geometry associated with a high street extent. The number of elements in the uprn array may be more than the address_count which identifies a subset of the address records which map to certain land use categories.			
toid	Array - Character	GeoJSON and PostgreSQL	
An array of the distinct Topographic Identifiers (TOIDs) identifying all of the polygon geometries classified as a building in the TopographicArea feature type of the product OS MasterMap Topography Layer which spatially intersect an address point geometry associated with a high street extent.			
usrn	Array - Numeric	GeoJSON and PostgreSQL	
An array of the distinct Unique Street Reference Numbers (USRNs) identifying all of the street records which the address point geometries associated with a high street extent are assigned to.			
postcode	Array - Character	GeoJSON and PostgreSQL	
An array of the distinct postcodes which the address point geometries associated with a high street extent are assigned to.			
address_geom	Geometry	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	
Multipoint geometry collecting all of the address point geometries associated with a high street extent.			
building_geom	Geometry	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	
Multipolygon geometry collecting all of the building polygon geometries associated with a high street extent.			
bbox_geom	Geometry	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	
Axis aligned bounding box or minimum bounding rectangle around the building_geom for a high street extent.			

Field Name	Data Type	Availability	ESRI Shapefile Field Name
orientated_bbox_geom	Geometry	GeoPackage, GeoJSON, ESRI Shapefile, and PostgreSQL	

A true, non-axis aligned bounding box or minimum bounding rectangle around the building_geom for a high street extent.

6. Notes

6.1 MHCLG Land Use Categories

The MHCLG land use categories are used annually to report on the amount and type of land changing use in England (https://www.gov.uk/government/collections/land-use-change-statistics). In most cases, the address classifications defined by a contributing local authority in the product OS AddressBase Plus can be mapped to one of the MHCLG land use categories in the table below. This mapping facilitates the aggregation of address geometries on a category-by-category basis to yield the address count attributes associated with each high street extent.

Sourced from p17 of the MHCLG Land Use Change Statistics in England: 2016-17 publication https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/712316/Land_use_change_statistics_England_2016-17.pdf

Table BN (Background Note) 1: Land use categories, groups and divisions

Previously developed land

Group	Category (codes)	
Residential	Residential	(R)
	Institutional and communal Accommodation	(Q)
Transport and Utilities	Highways and road transport	(H)
	Transport (other)	(T)
	Utilities	(U)
Industry and Commerce	Industry	(I)
	Offices	(J)
	Retailing	(K)
	Storage and warehousing	(S)
Community Services	Community buildings	(C)
	Leisure and recreational Buildings	(L)
Vacant	Vacant land previously Developed	(V - PDL)
Minerals and landfill	Minerals	(M)
	Landfill waste disposal	(Y)
Defence	Defence	(D)
Other developed use	Unidentified building	(~B)
	Unidentified general manmade surface	(~M)
	Unidentified structure	(~S)

Non-previously developed land

Group	Category (codes)	
Agriculture	Agricultural land	(A)
	Agricultural buildings	(B)
Forestry, open land and water	Forestry and woodland	(F)
	Rough grassland and Bracken	(G)
	Natural and semi-natural Land	(N)
	Water	(W)
Outdoor recreation	Outdoor recreation	(O)
Vacant	Vacant land not previously developed	(V - NPDL)
Residential Gardens	Residential Gardens	(RG)
Undeveloped land	Undeveloped land in urban areas	(X)

Ordnance Survey Limited
Explorer House, Adanac Drive, Nursling, Southampton SO16 0AS
+44 (0)3456 050505 (General enquiries) +44 (0)2380 056146 (Textphone)
customerservices@os.uk www.os.uk