DATASET FOR URBAN SCALE BUILDING STOCK MODELLING: IDENTIFICATION AND REVIEW OF POTENTIAL DATA COLLECTION APPROACHES

W. Y. Pei [1, 2], F. Biljecki [1, 3], R. Stouffs [1] *

[1] Department of Architecture, National University of Singapore, 117566 Singapore - stouffs@nus.edu.sg
[2] Future Cities Laboratory, Singapore-ETH Centre, 138602 Singapore - peiwanyu@u.nus.edu
[3] Department of Real Estate, National University of Singapore, 119245 Singapore - filip@nus.edu.sg

Commission IV, WG IV/9

KEY WORDS: City Dataset, Bottom-up Model, Urban Information Modelling, Building Material Stock, Data Standards.

ABSTRACT:

Construction materials play an important role in environmental impacts and make cities big resource consumers. To assess the sustainability of cities, the combined use of Life Cycle Assessment (LCA) and Material Flow Analysis (MFA) is considered effective to analyze construction material stock and flows. However, exhaustive data is required for such analyses, making LCA and MFA difficult to apply at the urban scale. Building information, the essential ingredient, is rarely available openly. Common approaches to gather the required data include both obtaining it directly from available datasets, e.g. open data from official sources, and indirectly generating data based on available data, e.g. using machine learning to fill the missing gaps. This research develops a data collection guideline for buildings' geometrical features, components and materials at the urban scale in the context of LCA and MFA. First, it identifies the basic steps of urban-scale building stock modelling and the list of data requirements. Second, the factors influencing the data collection are pointed out. In line with these guidelines, this research picks Singapore as a study area, reviewing the relevant authoritative open data sources and methodologies to estimate missing data. Finally, the suggestion on implementation of data collection are provided. When the data collection for urban scale stock modelling is limited by uncertain reality conditions, identifying and combining open datasets and data generation methods for data preparation is a necessity.

1. INTRODUCTION

In recent years, the estimation of construction material stock and flows at different geographical scales to analyze their environmental impact and achieve the use of secondary resources has gradually gained attention. To support policymakers and planners, it is indispensable to be able to predict and analyze urban material needs, as well as the evolving dynamics of the building stock. Most existing studies apply material intensity (MI) to estimate the total material stock that each study area has. However, building materials are often presented in complex assemblies of a material mixture integrated within components and structural systems, being difficult to recover directly. Even though quantifying aggregated building stock can help find the way of secondary material utilisation, there is still a need to estimate materials' specific forms (nature and physical state) and reconstruction technologies. To tackle this issue, more detailed stock modelling methods which combine geographic scales with building/component scales are needed for future urban materials circularity.

Existing methods to model the building material stock mainly include top-down and bottom-up approaches (Heeren and Hellweg, 2019). Top-down approaches collect data and information from historical macro-economic statistics, which do not explicitly consider individual physical factors specific to each type of building. Using a bottom-up approach, indicators for standard building material composition are defined, and a measure such as floor area is used to create a stock model for all buildings. Hence, bottom-up methods are typically used to study construction material flows and stock, allowing to trace all flows, from product, construction, and use to end of life and recovery stages.

* Corresponding author

A bottom-up stock model with building geo-referenced information can support MFA and LCA studies at a more detailed level. However, at a large scale, developing a bottom-up model requires sufficient information to connect building characteristics with material data, and modellers always have difficulty obtaining sufficient and accurate data, even when considering data from local and transnational data providers. For instance, building type and year of construction, which are essential for analyzing and distributing material-saving potentials in future scenarios for the city, are not available in many countries because of privacy issues and lack of data. A data-lite and proxy approach for stock modelling at the entire city scale consists of collecting quantitative information on building stock characteristics to identify building archetypes. However, a lack of accurate data for specific archetypes can cause significant uncertainties. Monteiro et al. (2018) noted that a lack of urban building data is a hindrance to research progress and is rarely addressed comprehensively in literature.

One approach to address these shortcomings is combining statistical survey data with archetype databases developed. In addition, improving the availability of geo-referenced data and using GIS are other approaches that can optimize the archetype bottom-up approach by adding time and space dimensions to the stock model and visualizing the material analysis results. These approaches and model outcomes are all influenced by data collection and input data. The existing MFA and LCA studies usually focus on specific cities/regions. However, the sources and approaches of one city to collect data might be different from another city, considering the available data sets are dissimilar. In this way, it is difficult to replicate the data collection approaches and compare the outcomes of MFA and LCA studies. Recently, the importance of tackling data issues and

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W2-2022-225-2022

developing data collection guidelines are realized by some researchers (Goy et al., 2020). To the best of the authors' knowledge, no studies have discussed the collecting data issue in support of building stock modelling on an urban scale.

This paper investigates the data collection of building material stock with geo-spatial characterization at the urban level. It reviews and evaluates both available datasets and data generation methods, emphasizing the need to conduct a comprehensive data collection for the different steps of modelling material stock. As such study is at the urban-level, it selects Singapore, an import-dependent country, as study area to adopt the data collection guidelines we outlined. This paper is structured as follows. Section 2 describes an overview of existing work. Section 3 summarizes the basic workflow of developing the large-scale material stock model and points out the data required. Section 4 reviews the data sources and data generation methods that can be applied for Singapore following the guidelines. Section 5 discusses the current data situation and data challenges of the case study, and section 6 concludes this paper.

2. LITERATURE REVIEW

It is a societal challenge to reduce the environmental impact of material consumption caused by infrastructure (e.g. buildings). As such, there is an inherent need to develop research approaches for transforming traditional material management patterns into sustainable and manageable ones. An increasing emphasis on analyzing material systems at the urban scale can help data management, visualization and, ultimately, developing a dynamic, spatial material stock model to improve the performance of simulating and assessing material flows. In recent years, numerous researchers have demonstrated that spatial proximity affects the analysis of urban materials' stock and flows (Augiseau and Barles, 2017). Almost all types of flows, stocks, trades, events, processes, distribution, lifetime and phenomena that researchers seek to explain in building material systems must occur in specific geographical locations. To map the building material system with geographical information, 3D city models are gradually being used as a visualization method to spatially represent both the natural and built/artificial features on a 3D scale (Khayyal et al., 2022).

Understanding cities as complex systems, a 3D model for a sustainable urban material system depends on reliable high-resolution data. Data acquisition technologies such as airborne imaging using UAVs and Light Detection and Ranging (LiDAR) can provide the necessary data for constructing urban 3D models. However, generating 3D models from remote sensing such as aerial or satellite imagery can be a costly, time-consuming and labor-intensive process. Besides, aerial or satellite imagery sources are not generally open to the public. As Chen et al. (2019) mentioned, more cities are moving to making open data available and making their use more efficient to support cities' material reuse and environmental goals. However, city models in 3D are more likely to be available in developed countries with higher economic levels or countries with national mapping agencies (Augiseau and Barles, 2017). In contrast, national mapping agencies and available resources in most developing countries for modellers to produce a 3D model are few. Nevertheless, there is a dearth of free high-level of detail 3D city models are available for use in many cities (Girindran et al., 2020).

Current studies show that different data resources, in combination, have the information content and geographical encoding potential to produce a single spatial material stock model. For example, Heeren and Hellweg (2019) collected data on all Swiss residential buildings from two national databases to form a 3D representation to derive the surface of construction elements and calculate the material stock. Mastrucci et al. (2017) collected geo-referenced footprints and attached attributes of Luxembourg buildings, such as building age and type, and airborne LiDAR data from available geo-spatial datasets provided by the municipality. Evans et al. (2017) described the British building stock using a 3D model called '3DStock', collecting data from two existing national datasets. Buffat et al. (2017) acquired 89% of footprints data from a cadastral survey and used OpenStreetMap (7%) and the Swiss cartographic SwissTLM dataset (4%) to fill data gaps. They further used Digital Surface Model (DSM) and Digital Terrain Model (DTM) raster datasets from the Swiss Federal Office of Topography to generate building heights and obtained building characteristics data from the Federal Register of Buildings and Dwellings.

In summary, it appears that most existing studies in this domain collect data from official government datasets. The main issue with this is a reliance on the availability of complete data sources for building information, which in reality does not exist in many urban areas. Urban material stock and flows analysis is therefore restricted in its widespread application by a lack of readily available datasets, as well as the labour-intensive processes required to produce those datasets (e.g. 3D city models or LiDAR data). Building documentation and resources are not available in most municipalities to develop such an effort from scratch. Within this context, some studies try to develop effective, data-lite modelling workflows adapted to current urban data structures or use open data sources to generate required data using simplified methods. For instance, Deng et al. (2022) developed urban stock modelling workflows using 'building archetypes' to represent a group of similar buildings. Others attempted to make up the data gap by generating data themselves. Deng et al. (2021) determined building types and formed the material stock model by integrating point-of-interest (POI) and community boundary datasets. Wurm et al. (2021) applied deep learning to generate data from aerial images and conduct building stock inventorying at a city scale.

For Singapore, the existing building stock can be an essential source of secondary materials. Some previous studies (Arora et al., 2019) quantified the materials and components of public residential buildings constructed by the Housing & Development Board (HDB) in Singapore. HDB is a government agency that ensures housing for all Singaporeans. The data was collected from HDB Property Information (https://data.gov.sg/), a tabular dataset containing a score of attributes for each building managed by HDB. However, for other (non-residential and private residential) buildings, it is much more challenging to obtain the required data comprehensively. Hence, accurate estimations of the material and building component stock in Singapore are still lacking. Therefore, after introducing the data collection guidelines, this study takes Singapore as a case study to identify datasets that can be directly collected and how other data can be estimated or generated using big data technologies.

3. DEVELOPING A GEO-REFERENCED BUILDING STOCK MODEL AT THE URBAN SCALE

3.1 Data Methodology

This study sets out to examine the data required for city-level, geographically specific building material stock modelling, ap-

plying a bottom-up approach combined with GIS. A disaggregated data collection is necessary to form a spatial-temporal GIS database integrating building data from different sources. Based on existing research applying GIS technology and combining a bottom-up method and spatial analysis to analyze building material stock at a large scale, the methodology includes three basic steps: 1) geo-spatial processing to improve the data structure; 2) spatialization of building material data; 3) establishing a dynamic and spatial model to simulate material stock and flow.

3.1.1 Step 1: Geo-spatial Processing to Improve the Data Structure Geo-spatial processing is the conversion of location-agnostic data structures to data explicitly associated with locations on earth. This step aims to arrive at an urban-scale GIS dataset that includes information on each building. For instance, buildings can be recorded as polygons with geographic locations. Building attributes also need to be recorded, including any building characteristics that relate to the calculation of materials, such as footprints, height, structural system, type, age, material, etc.

Upon preparing the datasets, the next step is to form the 3D urban model for mapping the building material information spatially. Using CityGML is the most straightforward approach to represent cities and urban areas for storing and exchanging geometrical 3D data of individual buildings (Wang et al., 2021). A Level of Detail (LoD) ranging from 0 to 3 is considered when describing datasets in CityGML, with geometric accuracy increasing with the level of detail. Models at LoD1 to LoD3 include three different types of elements, respectively: a box shape, sloped roofs, and texture on the exterior (such as windows and doors). The LoD for large-scale building material models typically ranges from 1 to 2 (Goy et al., 2020). The resulting 3D map of the urban area should be updated regularly, following any update of the data.

3.1.2 Step 2: Spatialization of Building Material Data Traditionally, many building and property-related datasets, especially with respect to building material information, have existed only in document format, not associated with any geographical data. Estimating the type and quantity of materials in the building stock is essential to successfully assess the environmental impact of construction material systems on a city level. In order to achieve this objective, a common approach is to identify a series building archetypes according to building functions and periods of construction and develop an archetype database. Specifically, the elements/components of reference building (such as walls, floors, roofs and windows) (see Figure 1) can be determined for each building archetype group. Then, a building's function and year of construction can be used to match to an archetype building with corresponding material-intensity coefficients. In this way, the layers of materials, composition, and thickness of every building element/component can be identified. After developing the fine-scale archetypes, which stand for different building groups, the results could be extrapolated to a larger level (region/urban) using the up-scaling factors such as the number of buildings per type or the floor area per type.

The synthesis of archetypes mainly involves two steps: segmentation and characterization. For the segmentation, the investigated building stock will be separated into categories based on building age and type. For the characterization, it will be necessary to determine specific material intensities per cubage and assign material properties to different archetypes representing the previously defined categories, either by using a real building sample or an 'average' virtual building based on statistical data.

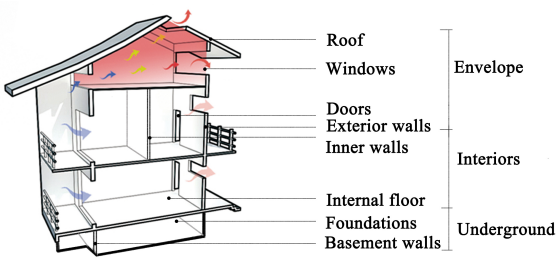

Figure 1. Building elements.

To quantify the spatio-temporal patterns of material stock at the urban level, it is necessary to analyze the unique geographic location of each building and map these to the reference archetype groups. In this way, the non-spatial material parameters can be assigned to individual buildings and analyzed for any specific urban area (e.g. considering administrative boundaries) using a clustering algorithm. Visualizing the results assists in identifying the presence and concentration of main construction materials within the city. This process is crucial to the accuracy and flexibility of the 3D material model, which in turn depends on the amount of data available in terms of both building structures and usage data.

3.1.3 Step 3: Establishing a Dynamic and Spatial Model to Simulate Material Stock and Flows Once the geo-referenced massing models and building archetype data are obtained, these can then be combined into an urban material simulation model. The final goal is to establish a dynamic and spatially explicit model that can represent and simulate the building material stock and flows at a city, regional, country, or even global level. Hereto, one needs to identify the end-of-life scenarios (demolition, transport, processing and disposal) for every material and building element/component. For example, the environmental impact of material flows during whole-building life cycle stages such as construction, retrofit, refurbishment, renovation, demolition and reuse of buildings should be identified and manifested. In principle, considering the dimension of time, material flow or inventory models can be dynamically built for a single year or longer time-scale, retrospectively (Gontia et al., 2020) or prospectively. However, the determination of end-of-life scenarios is still largely data-dependent, and time scale, construction, and flow rates can reach different accuracy levels depending on data availability. Quality and coverage of data vary considerably from country to country and even within a single city.

Spatial dynamics is another essential consideration in analysing material stock and flows systems. For the spatial dimension, developing a model combining Material Flow Analysis (MFA) and GIS is considered an effective approach. Such model can be used to analyze the spatial distribution and stock density (derived in relation to land area) and visualize the spatial heterogeneity of construction material stocks. A flow-driven urban building material model and spatially explicit MFA model ideally can describe material flows from one place to another (e.g. transporting construction waste from a demolishing building to a landfill). However, current research has not yet proposed any effective method to resolve the mismatch accrued during these flows or processes, such as when transport is con-

sidered on a daily or monthly scale, while the estimation of demolition rate is compiled on an annual or decadal scale.

In conclusion, spatial data achieved by transforming non-spatial information into geo-referenced maps forms the basic advantage of integrating building materials with spatial analysis. In addition, spatio-temporal patterns, trace material flow processes, and sources and hotspots of building material systems can be quantified based on a dynamic and spatial model to understand the drivers of the urban material system.

3.2 Data Needs

The data necessary for material stock modelling (including Life Cycle Analysis (LCA)) can be categorized as either geometrical features or as material data of buildings. This section describes the required data and their essential data characteristics.

3.2.1 Building Geometrical Features Data The geometrical features of a building dataset mainly include geo-referenced building footprints, building elevation data, and building characteristics such as year of construction and building type (see Table 1).

Content	Potential sources/method
Footprint/floor	OSM databases and official datasets
Building height	Datasets/DSM and DTM
Attributes (age/type)	Statistical source
Building gross volume	Computed: height × footprint
Area of walls	Computed

Table 1. Building geometrical features data.

Building footprints Building footprints can be used to indicate the boundaries of a building, associate other spatial datasets such as building materials data and estimate building dimensions. Building footprints should be provided as geo-referenced vector polygons corresponding to individual buildings. The quality of building footprint data is essential and building footprints from a cadastral survey generally have the best quality, as they are measured in the field by professionals. Many cadastral authorities maintain individual building footprints, as well as building attributes, such as their address, owner or construction date. However, a cadaster may not yet contain a digital record of every building's footprint, and availability may be restricted by data privacy and relevant data policies. Alternatively, open datasets such as OpenStreetMap (OSM) data can be used to obtain building footprints. Other methods can also be used to acquire building footprints, such as LiDAR and oblique photogrammetry, etc.

Building height The building height is an essential parameter to form a 3D building material model. Combined with building footprints, building heights can be used to virtually extrude a shoe-box model of buildings. Multiplied by lengths, building heights serve to obtain wall surface areas, which are related to the calculation of material mass. The number of floors can also be obtained, by dividing the building height by the average height of a single floor. Height data can be derived from analysing airborne LiDAR data and a DTM. A DSM and Digital Elevation Model (DEM) can also potentially be used to extract building heights. The height of residential and commercial buildings can also be calculated through multiplying the number of storeys by the floor-to-floor height, which is relatively fixed. The floor height for residential and commercial buildings may be assumed to be 3 m and 4 m, respectively (Deng et al., 2022). Some open 2D spatial datasets, such as OSM, also contain some height data. However, in many countries and areas, the third dimension is poorly represented in such datasets (less than a few percents of the buildings have height information in OSM) (Milojevic-Dupont et al., 2020; Biljecki, 2020).

Building characteristics Building attributes required for a material stock model typically include building type and construction year. Building type determines the use of the building, its layout (one floor or more), and its proximity to neighbouring properties (e.g. detached, semi-detached, terraced houses). Construction year can be used to identify the updates of building regulation codes, changes in the structural system, material types and construction technologies.

These attributes can be used to generate building archetypes. In most research, these attributes are collected from national/regional building attributes library datasets. If no building libraries are available, these may be inferred from other relevant datasets. For example, some housing websites contain considerable information relating to specific building types, such as year of construction, address, rental prices, etc.

3.2.2 Building Elements and Materials Data Compared with building geometry data, non-geometric data such as elements/components and material data (see Table 2) are relatively difficult to collect. In a large city, where there are thousands of buildings, it is almost impractical to collect non-geometric data for all of them. Therefore, it is necessary to develop building archetypes to represent different groups of similar buildings throughout the study area, based on building type and period of construction. When information about materials, building elements and their state of renovation is not available for individual buildings, these can be assigned to buildings in the stock according to the archetypes they match. Aside from data on building elements/components and data on building materials, other non-geometric data considered are data prepared for LCA.

Content	Potential sources/method
Material types and quantity	1) Divide building stock by building type and age 2) Identify reference building components
Material intensity/density	3) Identify characteristics of building components 4) Determine component's layers and thickness
Mass of each component	Computed (volume×intensity)
Service life of components	Lifetime modelling
Lifetime of materials	Statistical or reference data
Material flow of materials	Computed
Material treatment data	Statistical or reference data

Table 2. Building elements and materials data.

Data about building elements/components After segmenting the buildings in the study area based on building type and construction period to obtain the archetype buildings, information on building elements/components should be collected to describe these archetypes. This information relates to the structural system, components' amounts, types, and characteristics. Buildings of different periods generally have different structural systems and construction techniques adopted, which determine the type and amount of building components.

Information on structural building systems from different periods can be extracted from an archive and literature database. When not available, research papers on relevant construction details can also serve as a source. To identify the information related to building components, especially for residential buildings, building libraries are commonly used in many countries and regions. Alternatively, building component (and material layer) information can be identified based on building libraries from other countries with similar construction practices, or from statistics, technical standards, regulations, and the expertise of experts. Main outer components, such as outer walls, and main inner components, such as floors and inner walls, should all be listed, as they contribute significantly to material usage in buildings.

Data about building materials The archetype's structural system is relevant to decide on material type (timber, steel, concrete etc.), quantity and the refurbishment approach (related to LCA study). Calculating material quantity at different levels requires material data with different levels of detail. At the material level, the material stock of buildings is usually calculated using MI of different materials. The MI indicators are related to the material types, building function and the year of construction. At the component level, more detailed material data such as material layers, respective densities and thicknesses will need to be identified for each building element/component. The data for such material information can come from diverse sources, including construction department publications and handbooks providing standardized classification information and sample material inventories and characterizations.

Using all the information collected on building components, the mass of a particular material can be calculated from components' volume, area, or number, as well as its density or weight. For example, the material mass of a particular building category can be obtained by multiplying the gross volume of this building category with the specific MI for this material and building category. The total material stock then results from adding up the material masses for all building categories and materials.

Data prepared for LCA Upon obtaining the material stock, it is necessary to define prospective scenarios for analyzing the future material flows dynamically and the environmental impacts they would cause. The information prepared for LCA mainly includes the sampled service life for building elements, the lifetime of different building materials, flows (mass per time) of materials and the data of all material treatment processes such as transportation. Transportation data corresponds with urban transport and energy consumption of related activities such as cars, buses, motorbikes, trucks, as well as air traffic, for the treatment process of raw materials extraction, processing and material import. With this information, researchers can assess the environmental impact of buildings and materials over their life cycle.

Potential data sources leading to the synthesis of the life-cycle inventory include national and regional governmental corporations, national statistics, technical and science reports, guidelines, and previous studies.

3.3 Definitions and Influencing Factors of Data Collection

Since a wide range of data is required to develop a material stock model at the urban scale, data collection is a complex process, joining data from different sources. Hence, it is necessary to identify and review any factors influencing this data collection. These factors mainly include availability, accessibility, and data quality.

3.3.1 Data Availability The availability of building data is usually defined as its existence in any format (e.g., paper or digital format). As Goy et al. (2020) proposed, data availability at the urban scale is influenced by four main factors:

1. Location (city/country): The level of engagement with respect to building monitoring by municipalities and governments varies from one country to another. In some countries, the law necessitates a detailed characterization of buildings at a local level.
2. Time resolution: Annual construction material consumption data is more accessible compared to hourly data. Hourly data involves more complex monitoring, processing, and storage efforts.
3. Level of aggregation (from a single building to a larger area): Databases are always described by different levels of aggregation. It is much more challenging to collect individual building characteristics for the whole city.
4. Features and building types: Geometrical building information is more commonly available because of the application of advanced technologies such as airborne LiDAR or photogrammetry during land registration. However, other building characteristics, such as age, type, and structural system, are typically not documented at an urban scale.

3.3.2 Data Accessibility Data accessibility refers to how easily data can be accessed and used for material stock modelling. Some datasets may be available but not freely accessible to the public. Moreover, some data may be available in a paper-based form (construction documents) which needs to be transformed into electronic data before developing the digital stock model.

In recent years, more and more administrative entities have started to make some simple building information available on websites or online databases (e.g. building footprints). However, they commonly do not include the building height, age, component materials, or structural features. Even when such detailed data is not accessible to the public, it is available from the design and construction phase, especially in the case that BIM modelling has been adopted for building design. Addressing data accessibility issues and sharing available data is critical to improving current data challenges.

3.3.3 Data Quality Data quality is another important consideration as it directly affects the data collection work and the results from the analysis of a material system. Firstly, detailed building data is usually obtained from diverse databases and correlated features might not be consistent. Inconsistent correlations may be caused by measurement errors when developing the database. One solution is to merge different building databases based on their geo-location and reconcile any data gaps.

Secondly, considering a lack of data availability or accessibility, some assumptions and simplifications may need to be made for the study area. For example, when the required measurement data are unavailable, models often need to be made with analogies based on building data from neighbouring countries. Although these data have been validated for accuracy when studying the adjacent countries, it is impossible to assess the data accuracy when applied to the selected study area.

The effectiveness of data is another critical factor affecting its accuracy. Due to a lack of regular surveys and updates on construction data, modellers often need to use older data, which may inaccurately describe the current state of the material stock. In conclusion, data quality should be assessed before data collection and application, and a more thorough data cleaning should be conducted to ensure the quality of the datasets prepared for modelling.

3.4 Data Collection Approaches

While there is an increase in availability of datasets from cities and countries, such datasets are often spread across different sources and available in various formats for varying spatial and temporal scales. Hence, conducting data collection and improving datasets require multiple approaches. Two main categories of approaches can be distinguished: first, to merge data from accessible datasets; second, to generate required data applying appropriate technologies to accessible data.

Accessible datasets can be distinguished as official data sources and alternative ones. Official datasets are owned and authorised by government agencies. Data from official datasets tend to be more accurate, however, official datasets are not always open to the public. Alternative data sources may include data from private projects, open-access datasets, academic research, volunteered geo-information and some survey data (see Figure 2). Modelling building information, typically requires integrating data from several sources into a single data model.

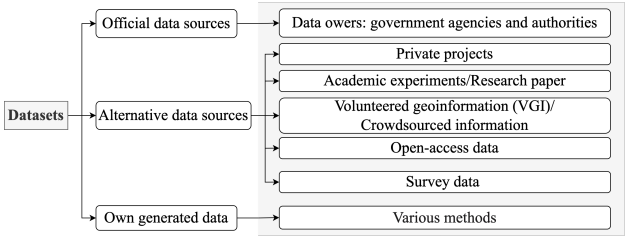

Figure 2. A classification of various data collection approaches

4. THE INVESTIGATION OF DATA SOURCES IN SINGAPORE

Based on the previous analysis of required data and factors influencing building stock modelling, this study adopts Singapore as an example to review potential datasets and data generation methodologies following the standard data collection guideline. Potential data sources include open-access datasets, academic research data and government datasets, etc. Quality assessment metrics are developed based on the influencing factors identified above.

4.1 The Development of Quality Assessment Metrics

4.1.1 Assessment metrics of open datasets For the assessment of open datasets, both data properties (data availability and accessibility) and data quality of the datasets are assessed. Data properties include 1) whether the dataset is free to download and use, or requires payment, and 2) the type of dataset reviewed (government/volunteer-driven data source, etc.). Data quality assessment metrics include:

1. Data scale and size: how large an area the data covers (city/district/neighborhood).
2. Data resolution: for example, pixels per inch for data in images format. For data in a 3D model format, the data resolution can be described using LoD.
3. Data accuracy: accuracy of geometric data mainly includes positional, shape (surface and edge), and orientation accuracy.
4. Data completeness: comprehensiveness or wholeness of the data, used to evaluate gaps or missing information.
5. Time (data age) accuracy: related to the effectiveness of data and frequency of data update.
6. Data formats: uniform data format and ability to remove formatting errors.

Among these metrics, time accuracy is one of the biggest challenges of data quality because the building sector and construction activities are dynamic. Due to the absence of data updates and regular surveys, the investigation of the timeliness of data updates mainly focuses on whether the dataset is updated regularly.

4.1.2 Assessment metrics of data generation methods The assessment of data generation methods mainly focuses on 1) inputs required to apply the method; 2) the professional knowledge required; 3) whether easy to update or not; 4) the limitations of the method. These metrics can be used to identify the usability of data generation methods when targeting different study areas. For example, some methods require input data not available in many countries and as such cannot be used. In addition, the expertise needed to use the method determines how difficult it is for modellers to apply this method.

As for assessing the quality of the data generated by different approaches, at this stage, this study mainly reviews the qualitative and quantitative evaluation of data generation provided by the existing research that proposed the approach. Some metrics are typically used to evaluate the generated data. For instance, the extraction results of building footprints or rooftops data are usually evaluated using intersection over Union (IoU), mean IoU (mIoU), Precision, Recall, F1-score, accuracy and Frame Per Second (FPS). However, since the proposed data generation approaches are applied to various study areas, verifying the data quality generated by different methods is difficult. This study considered to identify the available inputs and apply potential methods to the same urban area (Singapore) and compare the data generation results using the a serious of metrics.

4.2 Review and Assessment of Datasets

This study first reviewed the open data sources that can be used to collect building geometry and material data (link to summary table 3). In summary, OSM offers building footprint data for Singapore, however, height data is not included except for some specific regions in the city centre. Also, no research has yet assessed the quality (accuracy) of this height data. One map (OneMap@sla.gov.sg) is an authoritative national map in Singapore but is only available in image format, not in vector format.

Two other potential geometric data sources are Virtual Singapore and the 3D Singapore Sandbox. Virtual Singapore is an ongoing project championed by the National Research Foundation (NRF) to map Singapore in 3D. It is intended to be an authoritative 3D digital platform for use by the public, private companies, people and researchers. The 3D Singapore Sandbox developed by Singapore Land Authority (SLA) currently

provides users access to SLA's 3D geospatial data, including 3D models of over 160,000 buildings in Singapore. However, researchers cannot download the data or output data results even though they are encouraged to use the 3D geospatial data to develop and test new applications and services.

4.3 Review and Assessment of Data Generation Methods

Geometric and non-geometric building information is rarely available at an urban scale especially for developing countries. When some data is unavailable, it is essential to fix the data gap by generating the missing data. Hence, this study also reviews and assesses potential data generation methods (link to summary table 4). When there are multiple data generation methods available to generate the same data based on the same input data, a typical method was chosen for review and assessment.

The raw material for generating building footprints data mainly includes satellite images and drone data, enabling features to be digitized from high-resolution imagery. Modellers can apply deep learning technology such as a convolutional neural network (CNN) to automatically extract building features from remote sensing images. In recent years, automated feature extraction has made significant progress, especially when combining GIS with deep learning. This can benefit developing countries that do not have access to current data or the budget to get an accurate real-life map. However, the accuracy of using deep learning to extract building footprints is still questionable. As a result, some researchers combine the object detection task, which extracts feature values along with bounding boxes, and the semantic segmentation task, which classifies each pixel according to its properties.

For the generation of building height data, LiDAR allows high accuracy measurements but is not always available. Girindran et al. (2020) propose to combine the free version (low resolution) global Digital Surface Model (DSM), and Digital Elevation Model (DEM). Other methods that can be used to generate building height data also exist. In summary, building height information for Singapore may be generated using: 1) The number of storeys in a building; 2) Singapore regulations (e.g. FAR) and building footprints; 3) Shadows in high-resolution imagery; 4) Deep learning in combination with other building/city attributes such as footprints, facades and urban form.

For other required data, such as the year of construction, prior research has combined deep learning with map data, LiDAR data, or street view images. However, the accuracy of a data-driven building age estimation model will never be perfect. An as-built survey with extensive manual work is the only accurate way to assign a correct construction date to all buildings. Moreover, the combination of street-view images and a mobile-sensing approach is considered helpful in collecting information on building components and materials. Using data collected by mobile-sensing on target buildings within a district or urban area, modellers can develop the archetype database to build the material stock model at an urban scale.

5. DISCUSSION

This study summarized data requirements for material stock modelling and provided a guideline for the data collection step of the 3D material stock modelling. This model can be used to quantify the urban construction material stock, describe building stock development over time, point out the hot-spot material with high environmental impact and analyze the effectiveness of recycling strategies. Through reviewing available datasets and potential data generation methods, this study found that most available urban data in Singapore is 2D. For example, building footprints can be collected from OSM; Chen (2020) assessed the quality of Singapore residential building data in OSM and found that 97.67% of public residential building footprints are mapped in OSM. However, OSM data for Singapore does not include building heights, except for a specific region in the city centre. Even though OSMbuildings, an open source 3D map viewer, generates a 3D model for Singapore, a default height value is used for the extrusion when no specific data is available. While Virtual Singapore's 3D mapping of Singapore includes building heights, it is still unclear whether the data will be shared with the public in the future.

Collecting building data on residential buildings is still relatively feasible in Singapore. HDB flats contribute about 75% to the residential building stock, with the remaining made up of condominiums and landed properties. Property information such as maximum floor level and year of construction of HDB flats can be obtained directly from the Open Government Data Portal, Singapore (data.gov.sg/). HDB Housing unit drawings and layouts can be used to identify the building components. For the collection of other, non-residential buildings' geospatial data, archetypes could be developed. This methodology provides a path to classify buildings by usage categories and to determine the relevant building parameters for LCA studies (Buschka et al., 2021). Building types and years of construction of non-residential buildings can either be manually collected from various websites (e.g., URA SPACE) or generated using some available inputs such as street view images and historical satellite imagery. Thoma et al. (2014) developed a method for the classification of building types and periods leading to the material composition of buildings in Zurich, Switzerland. A similar method can be applied to Singapore based on knowledge acquired from building history research, old newspapers (NewspaperSG) and archives (National Archives of Singapore). In addition, combining open-access Digital Surface Model (DSM), Digital Elevation Model (DEM), and 2D building footprints is a potentially useful method to generate building height data in Singapore.

Building material data, such as the MI factors, may be collected from the Singapore Building and Construction Authority (BCA). For example, concrete is the most used construction material in Singapore, and volume of concrete can be computed by multiplying appropriate MI coefficients with accumulated usable floor areas. It is worth noting that MI coefficients relate to building type and year of construction. National concrete and steel consumption data can also be obtained from the Singapore Ministry of Trade and Industry (MTI). The building material flows data refer to the number of materials consumed by the construction industry and should be tracked within a certain period of time. New constructions and refurbishment are considered material inflows, while buildings demolished/refurbishment are treated as material outflows. For HDB flats, HDB annual reports provide demolition data, which can be used to calculate outflow of materials. Reports published by BCA also provide some information on the material stock of cement, steel, and aggregates. The number of materials consumed during any period of time/number of years, can be calculated using a raster technique from the difference between the respective datasets for the two different years.

6. CONCLUSION

This study outlines a data collection guideline for large-scale bottom-up building material stock modelling. Firstly, the workflow of material stock modelling is identified. After that, the data required for different modelling steps are pointed out, showing that a large amount of disaggregate data is needed for urban scale modelling. Hence, conducting a comprehensive data review to identify and integrate potential data sources is essential. Before reviewing the datasets, the factors influencing the input data quality are analyzed and metrics are developed for the assessment of these data sources. Facing the issue of hidden information at the urban scale, modellers first need to recognize which data can be collected from open datasets and which should be generated using technology approaches. Finally, this study reviews the most relevant open datasets and data generation methods for Singapore and provides suggestions or strategies for modellers. Specifically, government datasets and other open datasets are identified as the first choice for data collection because the quality of these datasets are controlled. Applying machine learning and some sensing approaches to generate required data is another useful approach for data collection. However, modellers need to have access to sufficient input data required to apply such approach and have the necessary professional experience. Future work will apply the recommendations from this study to the implementation of data collection and develop a high-resolution 3D building material stock model for Singapore.

ACKNOWLEDGEMENTS

The research was conducted partly at the Future Cities Lab Global at the Singapore-ETH Centre, which was established collaboratively between ETH Zurich and the National Research Foundation Singapore. This research is supported by the National Research Foundation Singapore (NRF) under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. The financial support provided by China Scholarship Council (CSC) to the first author is acknowledged.

REFERENCES

Arora, M., Raspall, F., Cheah, L., Silva, A., 2019. Residential building material stocks and component-level circularity: The case of Singapore. *Journal of Cleaner Production*, 216, 239–248.

Augiseau, V., Barles, S., 2017. Studying construction materials flows and stock: A review. *Resources, Conservation and Recycling*, 123, 153–164.

Biljecki, F., 2020. Exploration of open data in Southeast Asia to generate 3D building models. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, VI-4/W1-2020, 37–44.

Buffat, R., Froemelt, A., Heeren, N., Raubal, M., Hellweg, S., 2017. Big data GIS analysis for novel approaches in building stock modelling. *Applied Energy*, 208, 277–290.

Buschka, M., Bischof, J., Meier-Dotzler, C., Lang, W., 2021. Developing non-residential building stock archetypes for LCI—a German case study of office and administration buildings. *The International Journal of Life Cycle Assessment*, 26(9), 1735–1752.

Chen, E., 2020. Assessing the quality of openstreetmap building data in singapore. Master's thesis, National University of Singapore.

Chen, Y., Hong, T., Luo, X., Hooper, B., 2019. Development of city buildings dataset for urban building energy modeling. *Energy and Buildings*, 183, 252–265.

Deng, Z., Chen, Y., Pan, X., Peng, Z., Yang, J., 2021. Integrating GIS-based point of interest and community boundary datasets for urban building energy modeling. *Energies*, 14(4), 1049.

Deng, Z., Chen, Y., Yang, J., Chen, Z., 2022. Archetype identification and urban building energy modeling for city-scale buildings based on gis datasets. *Building Simulation*, Springer, 1–13.

Evans, S., Liddiard, R., Steadman, P., 2017. 3DStock: A new kind of three-dimensional model of the building stock of England and Wales, for use in energy analysis. *Environment and Planning B: Urban Analytics and City Science*, 44(2), 227–255.

Girindran, R., Boyd, D. S., Rosser, J., Vijayan, D., Long, G., Robinson, D., 2020. On the reliable generation of 3D city models from open data. *Urban Science*, 4(4), 47.

Gontia, P., Thuvander, L., Wallbaum, H., 2020. Spatiotemporal characteristics of residential material stocks and flows in urban, commuter, and rural settlements. *Journal of Cleaner Production*, 251, 119435.

Goy, S., Maréchal, F., Finn, D., 2020. Data for urban scale building energy modelling: Assessing impacts and overcoming availability challenges. *Energies*, 13(16), 4244.

Heeren, N., Hellweg, S., 2019. Tracking construction material over space and time: Prospective and geo-referenced modeling of building stocks and construction material flows. *Journal of industrial ecology*, 23(1), 253–267.

Khayyal, H. K., Zeidan, Z. M., Beshr, A. A., 2022. Creation and Spatial Analysis of 3D City Modeling based on GIS Data. *Civil Engineering Journal*, 8(1), 105–123.

Mastrucci, A., Marvuglia, A., Popovici, E., Leopold, U., Benetto, E., 2017. Geospatial characterization of building material stocks for the life cycle assessment of end-of-life scenarios at the urban scale. *Resources, Conservation and Recycling*, 123, 54–66.

Milojevic-Dupont, N., Hans, N., Kaack, L. H., Zumwald, M., Andrieux, F., de Barros Soares, D., Lohrey, S., Pichler, P.-P., Creutzig, F., 2020. Learning from urban form to predict building heights. *Plos one*, 15(12), e0242010.

Monteiro, C. S., Costa, C., Pina, A., Santos, M. Y., Ferrão, P., 2018. An urban building database (UBD) supporting a smart city information system. *Energy and Buildings*, 158, 244–260.

Thoma, E., Fonseca, J. A., Schlueter, A., 2014. Estimation of base-values for grey energy, primary energy, global warming potential (gwp 100a) and umweltbelastungspunkte (ubp 2006) for swiss constructions from before 1920 until today. *Contemporary Urban Issue Conference Informality: Rethinking the Urban (CUI 2014)*, ETH Zurich, Architecture and Building Systems.

Wang, C., Wei, S., Du, S., Zhuang, D., Li, Y., Shi, X., Jin, X., Zhou, X., 2021. A systematic method to develop three dimensional geometry models of buildings for urban building energy modeling. *Sustainable Cities and Society*, 71, 102998.

Wurm, M., Droin, A., Stark, T., Gei, C., Sulzer, W., Taubenbck, H., 2021. Deep learning-based generation of building stock data from remote sensing for urban heat demand modeling. *ISPRS International Journal of Geo-Information*, 10(1), 23.