Response to Reviewers

From Archives to AI: Residential Property Data Across Three Decades in Brunei Darussalam (Data in Brief)

Haziq Jamil Amira Barizah Noorosmawie
Hafeezul Waezz Rabu Lutfi Abdul Razak

2025-02-22

Reviewer #1: The manuscript presents a valuable dataset covering Brunei's residential property market from 1993 to 2024, which is likely to serve as an important resource for researchers in economics, urban planning, and real estate. I have some thoughts on improvement that could be implemented to highlight and make clear the limitations and applications of the data created.

Firstly, the data collection approach combines manual transcription from newspaper archives, web scraping, and LLM-based cleaning. While this multifaceted method is innovative, the reliance on manual processes and Excel for spatial harmonisation raises concerns about reproducibility and scalability. Wondering if the author has thought about this - not just in terms of introducing human error - but also how it can contribute the longevity of the data source moving forward.

Second, since the dataset is predominantly derived from newspaper listings and online property advertisements, there is a risk that certain segments of the market—particularly properties not advertised through these channels—may be under-represented. Incorporating additional sources, such as government records (for example, land registries or records of land transactions), along with validation from real estate agents, could contribute to a more representative sampling of the residential property market.

Thirdly, the LLM-based cleaning process is a notable innovation in this study. However, with an accuracy rate of 93%, there is potential for errors to propagate in the final dataset. A more detailed error analysis or cross-validation with alternative cleaning methods would enhance confidence in the dataset. In addition, the operational and computational costs associated with employing LLMs should be taken into account, as these may affect the scalability of the approach.

Finally, it is important to note that the dataset relies on listing prices as a proxy for market values. Listing prices do not always reflect actual sold prices due to negotiation dynamics and marketing strategies. A factual discussion of this limitation is recommended, particularly if market trends or policy implications are to be derived from the analysis.

Handling editor -

- The LLM-based data cleaning introduces subjectivity and potential bias. Some elaboration here is required.
- The paper claims significance for policy applications, but the dataset is specific to Brunei, limiting generalizability. some elaboration would be appreciated.
- How was data accuracy validated (especially for manually transcribed entries)? What are the error margins for web-scraped data?
- How consistent is the extraction across different property descriptions? LLM outputs may suffer from data hallucination—was there a manual verification step?
- The LLM model and parameters are not fully described (e.g., temperature setting, prompt structure). The dataset is not fully independent of subjective processing.
- and finally: focus on data curation, not economic analysis!

Scientific Editor -

1. Specifications table/data accessibility: Please make the data citable via zenodo and include a link here.

Reviewer's Responses to Questions

1) Are these data original and produced by the authors?

Please respond with Yes OR No OR N/A.

Reviewer #1: Yes

2) Are these data secondary (e.g. censuses, government databases, organizational records)?

Please respond with Yes OR No OR N/A. If YES, please answer 2a, 2b & 2c; if NO go to 3

Reviewer #1: Yes - though data has been sourced from existing records, such as newspaper advertisements and online property listings, not census or records.

2a) Secondary Data Only: were these data collected using variables that make the study unique?

Please respond with Yes OR No OR N/A.

Reviewer #1: Yes

2b) Secondary Data Only: is this collection of secondary data of value to the research community?

Please respond with Yes OR No OR N/A.

Reviewer #1: Yes

2c) Secondary Data Only: do the authors provide the protocol for collecting/creating these data?

Please respond with Yes OR No OR N/A.

Reviewer #1: Yes

3) Have the authors used a questionnaire or survey?

Please respond with Yes OR No OR N/A. If YES, please answer 3a; if NO go to 4.

Reviewer #1: No

3a) Is the sampling representative of the population and rigorously following a scientific method?

Please comment on the rigor of the sampling method and if additional sampling or a different sampling method is required. Please also mention if the questionnaire/survey being used is direct, unambiguous and unbiased.

Reviewer #1: The sampling method primarily relies on collecting property listings from newspaper archives and online platforms, which essentially constitutes a form of convenience sampling.

It is hard to tell if the sampling is biased to agents that actively list their properties, long-standing RE companies, and whether there are missing listings that make an impact to the overall usability and generalisability of the data produced.

If/wherever possible, it would be beneficial to incorporate additional data sources. Government records—such as those from land registries or records of land transactions—could provide an official baseline. Also, engaging with real estate agents for corroboration could help to identify any systematic biases in the listings and ensure that the data reflects actual market conditions, rather than merely the prices at which properties were initially advertised.

4) Do the authors adequately explain to the research community the utility of these data in the "Value of data" section?

Please include a comment on the validity of this section. Include notes on how this can be improved, if necessary.

Reviewer #1: Yes, though could have more meat on the bones.

5) Are these data described clearly in the "Data description" section?

Please provide suggestions to the author(s) on how to further clarify the presentation and description of the dataset.

Reviewer #1: Yes

6) Is the protocol/method for generating these data adequately described in "Experimental design, materials, and methods" section?

Please include suggestions on how the section can be improved to aid reproducibility/reusability.

Reviewer #1: Authors should consider the following:

Authors describe several methods for data collection, the reliance on manual transcription and Excel-based spatial harmonisation raises concerns regarding reproducibility and scalability. Perhaps we need to discuss this.

The integration of historical data is a notable strength, yet differences between manually collected early data and later web-scraped data may introduce inconsistencies. The manuscript would benefit from further discussion on how these differences might affect comparability over time.

The use of an LLM-based cleaning method is innovative; however, with a reported accuracy of 93%, some errors may persist. A more detailed error analysis or cross-validation with alternative cleaning methods would provide additional assurance regarding the data's accuracy.

Another important aspect to consider is the reliance on LLMs for data cleaning. The cost associated with deploying LLMs—both in terms of computational resources and the expertise required to maintain and update these systems—should not be underestimated. The financial and operational costs may impact the scalability of the workflow, particularly if ongoing adjustments are needed to accommodate changes in data sources or to improve model performance.

7) Have the authors provided all the raw data related to all the tables, graphs, images and charts, etc. and are they freely accessible?

Please provide suggestions to the author(s) on how to improve data accessibility for wider usage. Please mention missing raw data, if any.

Reviewer #1: Yes

8) If this data article is related to an existing primary research article is there any duplication? If yes, please comment on this.

Please mention any overlapping text, images, etc.

Reviewer #1: Not known

Reviewer 1

Reviewer 2

References