Centre for Population Ageing Research (CEPAR)

Working Paper 2025/01

Hierarchical House Price Model Incorporating Geographical and Macroeconomic Factors: Evidence from Australia

Lingfeng Lyu, Yang Shen, Michael Sherris, Jonathan Ziveyi

Hierarchical House Price Model Incorporating Geographical and Macroeconomic Factors: Evidence from Australia

Lingfeng Lyu[1,2], Yang Shen[1,2], Michael Sherris[1,2], Jonathan Ziveyi[1,2]

[1]School of Risk and Actuarial Studies, UNSW Sydney, NSW 2052, Australia.

[2]ARC Centre of Excellence in Population Ageing Research (CEPAR), UNSW Sydney, UNSW Sydney, NSW 2052, Australia.

Abstract

This paper proposes a tri-level hierarchical model for house price prediction at Australian suburbs postcode level, integrating dynamics from the national level and the Statistical Areas Level 4 (SA4) level under the Australian Statistical Geography Standard (ASGS). Our study advances house price modelling by introducing a novel framework that integrates risk premium–principal component analysis (RP-PCA), vector autoregressive (VAR) modelling, and an empirical copula approach. Employing RP-PCA to extract SA4-level risk factors and combining these with national-level drivers, we develop a VAR model to capture dynamic relationships. Spatial dependencies in one-step-ahead forecast residuals across suburbs are modelled via an empirical copula, further enhancing predictability. Results demonstrate that this geographically conditional multi-factor model, structured hierarchically, increases interpretability and improves short-term forecast accuracy without compromising long-term robustness. Furthermore, this methodology presents a dynamic and granular view of house price trends in Australia. Results highlight key national determinants, including interest rate shifts, gross domestic product growth, and exchange rate variations, particularly in metropolitan urban areas. At the SA4 level, household debt levels, income growth, and population dynamics emerge as critical determinants of price trends, highlighting the interplay of economic and demographic drivers across spatial scales.

Keywords: Hierarchical framework; House prices; Spatial dependencies; Risk premium - principal component analysis (RP-PCA).

1 Introduction

Home equity constitutes a major component of individual wealth. In Australia, the residential real estate market now exceeds AUD 10 trillion (Actuaries Institute, 2024).[1] Localised house prices affect economic behaviour (Atalay and Edwards, 2022; Qian, 2023), including retirement planning (Hanewald and Sherris, 2013; Shao et al., 2015), and this impact is particularly pronounced given the social insurance system's close connection to home equity.[2] Therefore, a detailed, spatially refined understanding of house prices is crucial for informing public policy, guiding financial decisions, and ensuring equitable service provision. Evidence from house price economics reinforces the importance of granular analysis in that modelling housing prices at a disaggregated level—categorised by submarket, time period, or property type—enhances predictive accuracy (Bin, 2004; Bourassa et al., 2003; Fève and Moura, 2024; Goodman and Thibodeau, 2003; Granziera and Kozicki, 2015; Minetti et al., 2019). In the Australian context, Chomik and Yan (2019) identify four core drivers of housing supply and demand, influencing house prices: (1) population growth (Gevorgyan, 2019; Mankiw and Weil, 1989; Takáts, 2012), which elevates demand through demographic shifts; (2) inflation and interest rates (Goodhart and Hofmann, 2008; Shi et al., 2014), shaped by central bank policies; (3) supply responsiveness (Caldera and Johansson, 2013; Hilber and Vermeulen, 2016; Saiz, 2010), which in Australia has historically lagged demand; and (4) income levels (Lee et al., 2022; Wetzstein, 2017), reflecting household purchasing power. Together, these factors illustrate the multi-layered forces shaping housing markets, yet their integration into predictive models remains fragmented.

A foundational stream of research emphasises macroeconomic drivers of house price dynamics such as interest rates, gross domestic product (GDP) growth, and monetary policy. For instance, Goodman and Thibodeau (2008) link speculative housing behaviour to macroeconomic cycles, while Kuttner and Shim (2012) demonstrate how credit conditions uniquely influence Australian house prices. Cross-country analyses in Adams and Füss (2010) further underscore the interdependence of housing markets and broader economic growth, suggesting global factors cannot be ignored. Moreover, Zhu et al. (2017) investigate the influence of monetary policy stance and mortgage market structure in European countries. These studies, however, often focus on national or international scales, without considering sub-national dynamics, as evidenced in recent spatio-temporal studies (Aquaro et al., 2021; Baltagi and Li, 2014; Brady, 2011; Shen and Pang, 2018).

Complementing this macroeconomic focus, recent work highlights localised risk factors such as labour markets, credit accessibility, and neighbourhood amenities. Green and Malpezzi (2001) tie housing demand to local employment trends, while Guerrieri et al. (2013) show how area-specific lending practices amplify price fluctuations. Glaeser and Gyourko (2018) attribute rising price differentials to the uneven distribution of urban amenities, a theme extended in Titman and Zhu (2024), linking house price dynamics to city-specific

[1] The residential real estate market is estimated at approximately USD 6.4 trillion, based on an exchange rate of 1 AUD being equivalent to 0.64 USD.

[2] Means-tests for the age pension and aged care in Australia account for homeownership or home equity values (My Aged Care, 2024a,b; Service Australia, 2024).

agglomeration externalities and land-use policies. In addition, Bhattacharjee et al. (2016) and Wrenn et al. (2017) demonstrate that local fundamentals—including issues of house price endogeneity and land supply constraints—play a pivotal role in shaping submarket dynamics, while evidence from Brady (2011), Basile et al. (2014), and Shen and Pang (2018) further highlights how shocks propagate across space and time at the local level. Similarly, Ganduri and Maturana (2024) introduce property rehabilitation activity as a micro-level explanation for neighbourhood price trajectories, challenging purely macroeconomic explanations

Despite considerable progress in modelling efforts at granular geographical levels, existing approaches frequently treat local drivers as static, isolated inputs rather than dynamic interacting components. Moreover, few studies integrate these dynamic and multi-level factors within a unified predictive framework. To address these limitations, we propose a tri-level hierarchical model which integrates macroeconomic risks (such as bond yields and GDP growth) with observable micro-level drivers (such as demographics, income, and housing stock). This hierarchical model is informed by top-down methodologies (Gross and Sohl, 1990; Hyndman et al., 2011), which offer two key advantages for our context. Firstly, this approach facilitates incorporating risk factors across geographical scales. Secondly, bottom-up approaches have been shown to produce suboptimal results for highly disaggregated datasets due to their tendency to neglect cross-series dependencies (Dunn et al., 1976; Hyndman et al., 2016; Orcutt et al., 1968).

This paper employs a geographically conditional capital asset pricing model (CAPM), adapted from Jagannathan and Wang (1996), to develop a hierarchical framework that integrates various levels of risk factors. To address overfitting, the model employs risk-premium principal component analysis (RP-PCA), an adjusted principal component analysis (PCA) with penalty terms (Lettau and Pelger, 2020). Then, we employ vector autoregressive (VAR) modelling to capture dynamic relationships of risk factors. Moreover, using information from a model's residual covariance structure enhances prediction accuracy, as evidenced in (Nystrup et al., 2020; Wickramasuriya et al., 2018). Therefore, to further enhance model accuracy, this paper describes the residuals of the primary house price model using a multi-variable empirical copula (Ho et al., 2016), demonstrating the conditional covariances across various areas resulting from missing information in data collection.

The results indicate that national-level variables significantly influence house prices, including interest rates, exchange rates, gross domestic product (GDP) growth, and retail sales. Higher interest rates raise borrowing costs, dampening housing demand, particularly in metropolitan areas such as Sydney and Melbourne. Exchange rate fluctuations affect foreign investment, influencing suburbs with substantial foreign investment, such as the central business areas of Sydney and Melbourne. GDP growth signals economic strength, directly driving house prices in economically pivotal areas. Similarly, retail sales growth reflects heightened market activity, contributing to rising house prices in commercial hubs.

The results also show that, at the Statistical Area Level 4 (SA4), factors like wage growth, financial distress, and home improvement activities affect house prices.[3] Wage growth boosts purchasing power

[3]The Australian Statistical Geography Standard (ASGS) is a social geography that categorises Australia into a hierarchy of

and raises house prices, especially in wage-driven suburbs, while financial distress lowers them. Home improvement often signals rising home equity values. Demographic factors also matter: Population growth pushes up house prices in densely populated areas due to limited supply, and areas with a wider age range and higher education levels tend to have higher house prices due to stronger purchasing power and demand. In Australia, the influx of immigrants further contributes to these trends by increasing urban populations, thereby intensifying competition for housing and sustaining upward price pressure.

This study advances postcode-level house price modelling by integrating a comprehensive set of variables to improve interpretability, building on foundational frameworks (Jagannathan and Wang, 1996). Unlike prior work that treats spatial risk factors as static or isolated (Hanewald and Sherris, 2013; Shao et al., 2015), we introduce novel area-varying risk premia and dynamic risk factors operating across multiple geographical scales. Our model also uniquely accommodates mixed-frequency data spanning these scales, addressing a critical gap in existing methodologies (Ghysels, 2016).

The second contribution lies in the framework's ability to enhance interpretability without compromising predictability. The proposed RP-PCA-VAR-copula model achieves higher short-term prediction accuracy and prevents significant deterioration in long-term forecasting performance compared to benchmark models such as the factor-augmented vector autoregressive (FAVAR) model (Bernanke et al., 2005) and the MinT(Shrink) approach (Wickramasuriya et al., 2018). Furthermore, the framework demonstrates the viability of extending the RP-PCA methodology (Lettau and Pelger, 2020) to spatial settings, thereby broadening its applicability beyond traditional asset pricing contexts to spatially correlated systems.

The remainder of the paper is structured as follows: Section 2 details the dataset and preprocessing steps used to prepare it for analysis. Section 3 introduces our hierarchical house price model and explains how risk factors are developed. Section 4 presents the results, validates the model's structure, tests its predictive accuracy, and analyses risk factors through sensitivity tests to demonstrate economic relevance. Finally, Section 5 discusses the implications of our findings, highlights limitations, and suggests directions for future research.

2 Data

CoreLogic is a leading provider of standardised data and services related to home equity across Australia (CoreLogic, 2024). The company offers monthly house price index (HPI) series for various areas at the postcode level, with data available since January 1980. As shown in Figure 1, these indices are pre-processed and standardised to a base value 100 in December 2009. This standardisation prevents direct comparison of absolute house price indices across areas but allows relative changes between areas to be derived, serving as

statistical areas based on the location of people and communities, designed by the Australian Bureau of Statistics (ABS) for statistical release and analysis, taking into account the requirements of statistical collections and relevant geographic concepts. The structures include six interrelated hierarchies of areas: Mesh Blocks (MBs), Statistical Areas Level 1 (SA1), Statistical Areas Level 2 (SA2), Statistical Areas Level 3 (SA3), Statistical Areas Level 4 (SA4), States and Territories (S/T), and Australia.

the basis for house price modelling.

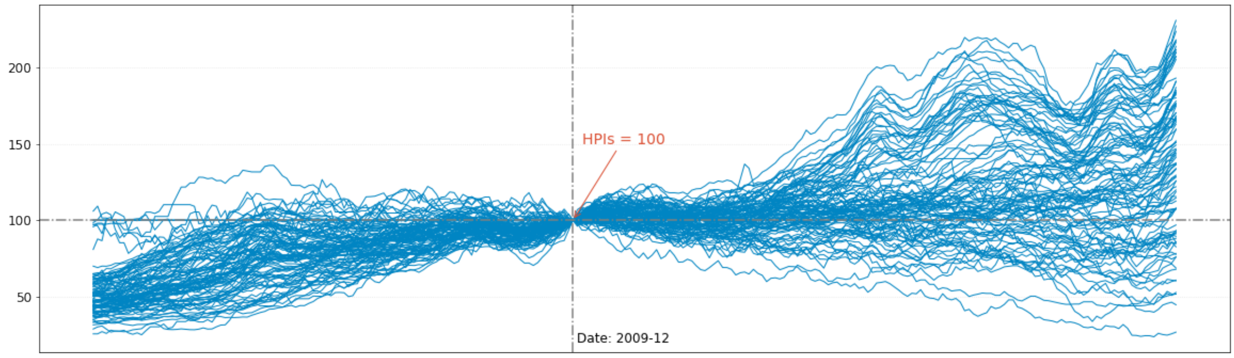


Figure 1. HPI series in all suburbs sourced from CoreLogic (January 2000 to December 2020). All series have been pre-processed such that HPIs are standardised to 100 in December 2009.

Table 1. General introduction of data at different geographical levels.

Level	**Variable name**	**Frequency**	**Source**
Postcode	Coordinates	NA	ABS
SA4	annual income, language, employment, working hours, education level, expenditure, wealth, debt, population size, population structure	annual	HILDA
National	interest rate, exchange rate, retail sales, private dwelling approvals, total dwellings, dwellings structure	monthly	RBA
	CPI, GDP	quarterly	RBA
	S&P/ASX 200	monthly	Yahoo Finance

Notes: This table summarises the data sources and frequencies for the series, excluding the house price index (HPI), across different geographical levels. Postcode-level coordinates are sourced from the Australian Bureau of Statistics (ABS) (Australian Bureau of Statistics, 2023), while SA4-level data is derived from the Household, Income and Labour Dynamics in Australia (HILDA) Survey. National-level data is obtained from the Reserve Bank of Australia (RBA) (Reserve Bank of Australia, 2023) and Yahoo Finance (Banasiak, 2016).

Common macroeconomic variables are collected to incorporate relevant risk factors influencing the housing market. Following Chomik and Yan (2019), a range of variables affecting area-specific housing supply and demand is included. Annual data from the Household, Income and Labour Dynamics in Australia (HILDA) Survey (Melbourne Institute, 2019) provides insights into demographics, income levels, and employment status across suburbs. Key population-related drivers of house prices, such as age, gender, family size, language, education, and employment, are identified. Moreover, wealth and debt are connected to the second driver: banking deregulation and credit access. The longitude and latitude data, often used in previous research to represent housing demand (Liu et al., 2016; Pace et al., 2000), serve as a proxy for housing supply in a given area.[4] Table 1 catalogues these variables, their frequencies and sources.

[4]While not directly included as variables in the model, these coordinates will be referenced in the numerical analysis to explore geographical variations in house prices.

As Table 1 indicates, only two national variables, CPI and GDP, have a quarterly frequency. These variables will be linearly extrapolated to generate monthly data. It is also important to note that the variables from HILDA only serve as a broad categorisation, representing classes in which all variables can be described. The exact variable names and the associated preprocessing methods will be discussed and shown in Table 2 subsequently.

Categorical variables are processed as proportions for each area at different levels. The total number of individuals in the HILDA survey is a parameter in the model, reflecting its representativeness of the broader Australian population. This inclusion accounts for such differences as the HILDA sample size varies across areas based on population characteristics (Melbourne Institute, 2019). After testing and addressing potential sources of non-stationarity, such as unit roots or seasonality, variables or their first- or second-order differences are incorporated into the variable list, following standard procedures for handling non-stationary time series (Lutkepohl and Kratzig, 2004).[5]

To maintain interpretability and reduce complexity, we limit the model to three levels, namely national level, SA4 level, and postcode level. Identifying a middle level is necessary to align the HILDA data with the CoreLogic data for each postcode. Since CoreLogic includes more postcodes than HILDA, a selection criterion for the middle level is established: For each CoreLogic postcode, a corresponding middle-level area in HILDA must exist. If a specific postcode is absent from HILDA, other postcodes within the same middle-level area must be present to derive risk factors for that level. The SA4 level is identified as the most granular level satisfying this criterion.[6] The annual SA4-level data is first treated as a single dataset and then extrapolated to a monthly frequency before being incorporated into the full hierarchical model.

3 A dynamic house price modelling framework with hierarchical structure and residual analysis

We assume HPIs in Australia are hierarchical and nested within different geographical jurisdictions. Our basic model setup consists of two parts. The first part is a three-level hierarchical model describing HPI growth rates, where an RP-PCA model is applied to find the risk factors at the SA4 level (Lettau and Pelger, 2020). Then, the second part incorporates an empirical copula into the model to analyse residuals from the hierarchical model.

3.1 Problem statement

We design a three-level hierarchical factor model to capture the underlying idiosyncratic demand for home equity. Data at different levels are exogenously collected, as shown in Section 2 to estimate these factors. We

[5] Due to the limited length of HILDA data, CoreLogic HPIs are truncated, and the final model uses data from the last 15 years (2006–2020), ensuring compatibility with preprocessed variables and accounting for the unavailability of certain HILDA variables before 2005.

[6] Only one SA4-level area, `White-Bay`, does not meet this criterion and is excluded from the analysis.

Table 2. Detailed description of variables and their stationary methods.

Variable name	Description	Stationary method
ir	Interest rate	1
exr	Exchange rate	1
cpi	Consumer price index	1
gdp	Gross domestic product	1
rs	Retail sales	1
pda	Private dwelling approvals	1
asx	Australian securities exchange	1
hsdebt	Total home debt	1
wsce	Current weekly gross wages & salary	1
agemedian	Median of age	0
agestd	Standard deviation of age	1
ageq3	The third quantile of age	1
agecount	Total number of people	1
xphmrna	Home repairs/renovations/maintenance	1
xphltpa	Fees paid to health practitioners	2
lsemp	Weekly time on paid employment	1
iprbwm	Went without meals	0
fiprbmr	Could not pay the mortgage/rent on time	1
ancob_1101	Country of birth: Australia	0
ancob_1201	Country of birth: New Zealand	1
ancob_7103	Country of birth: India	1
ancob_6101	Country of birth: China (excludes SARs)	1
ancob_5105	Country of birth: Vietnam	2
edhigh1_4	Highest education level: Year 12 or equivalent	0
edhigh1_5	Highest education level: Certificate III/IV	1
edhigh1_8	Highest education level: Bachelor degree	1
edhigh1_9	Highest education level: Postgraduate degree	1
anatsi_2	Aboriginal	2
anatsi_3	Torres Strait Islander	1
chkb12_1	Employment status	2

Notes: The information presented in the first and second sections pertains to numerical variables at national and SA4 levels, whereas the information in the last sections pertains to categorical variables at the SA4 level. For categorical variables, the data is processed to show the proportion of each group in a given area. Different numbers represent the order of differentiation.

assume $\tilde{I}_{i,t-1}$ represents the complete information set of the i^{th} suburb, encompassing all previous economic factors influencing house prices up to time t, though not directly accessible. To manage this, a hierarchical partition is introduced to represent accessible economic information, with varying information sets defined by different partitions of suburbs.

Definition 3.1. *The hierarchical partition is defined as:*

$$\begin{aligned} &\Omega_j \cap \Omega_{j'} = \emptyset, \ \textit{for } j \neq j'; \ \cup_j \Omega_j = \Omega, \ \textit{where } j \in \{1, ..., N\}; \\ &\Omega_{ji} \cap \Omega_{ji'} = \emptyset, \ \textit{for } i \neq i'; \ \cup_i \Omega_{ji} = \Omega_j, \ \textit{where } i \in \{1, ..., N_j\}. \end{aligned} \tag{1}$$

Here Ω is the universal set with all suburbs. Ω_j includes all suburbs belonging to area j at the SA4 level, where $j \in \{1, ..., N\}$, and N is the total number of SA4-level areas. Ω_{ji} is a single-element set that contains the i^{th} suburb belonging to area j at the SA4 level. Similarly, N_j is the total number of suburbs in area j ($i \in \{1, ..., N_j\}$). We assume that the CAPM holds in a conditional sense in the analysis of HPI growth rates.

Definition 3.2. *The accessible information is defined through specified partitions, with inclusion relationships as follows:*

$$I_{i,t-1}(\Omega_{ji}) \subset I_{i,t-1}(\Omega_j, \Omega_{ji}) \subset I_{i,t-1}(\Omega, \Omega_j, \Omega_{ji}) \subset \tilde{I}_{i,t-1}, \tag{2}$$

where $I_{i,t-1}(\Omega_{ji})$ represents the information set at the postcode level, $I_{i,t-1}(\Omega_j, \Omega_{ji})$ represents the information set at both the postcode and SA4 levels, and $I_{i,t-1}(\Omega, \Omega_j, \Omega_{ji})$ represents the information set at all levels.

Remark 3.1. *The information gap between the complete information set $\tilde{I}_{i,t-1}$ and the largest accessible information set $I_{i,t-1}(\Omega, \Omega_j, \Omega_{ji})$ contributes to model inaccuracies.*

Postcode-level excess HPI growth rates are treated analogously to excess asset returns in this paper, following the approach of Fama and MacBeth (1973). A multi-factor CAPM is utilised to estimate risk premia. Static CAPMs, which assume constant risk factors across periods and areas, fail to adequately capture underlying risks, making them overly restrictive for house price modelling. To address this limitation, the analysis assumes that CAPM holds conditionally when modelling HPI growth rates.

Lemma 3.1. *For the excess HPI growth rate of Suburb i in period t,*

$$\mathbb{E}\left[h_{i,t} \mid \tilde{I}_{i,t-1}\right] = \mathbb{E}\left[\mathbf{w}_i \mid \tilde{I}_{i,t-1}\right]^\top \mathbb{E}\left[\boldsymbol{f}_{i,t} \mid \tilde{I}_{i,t-1}\right], \tag{3}$$

where $\mathbf{w}_i$ is a vector of coefficients describing risk premia, and $\boldsymbol{f}_{i,t}$ is a vector of risk factors for the i^{th} suburb.[7]

To explain the variations in the unconditional expected excess HPI growth rate across suburbs, we first take the expectation of both sides of Equation (3) about $\tilde{I}_{i,t-1}$.

Lemma 3.2. *Under the assumption that each complete information set $\tilde{I}_{i,t-1}$ at time t is available, the unconditional expected excess HPI growth rate is:*

$$\mathbb{E}\left[h_{i,t}\right] = \mathbb{E}\left[\mathbf{w}_i\right]^\top \mathbb{E}\left[\boldsymbol{f}_{i,t}\right] + \mathbb{C}\text{ov}_\omega\left(\mathbb{E}_\omega\left[\mathbf{w}_i \mid I_{i,t-1}\right], \mathbb{E}\left[\boldsymbol{f}_{i,t} \mid I_{i,t-1}\right]\right). \tag{4}$$

As $\tilde{I}_{i,t-1}$ is defined as the complete information set of the i^{th} suburb, $\mathbb{E}_\omega\left[\mathbf{w}_i \mid \tilde{I}_{i,t-1}\right] = \mathbf{w}_i$, and $\mathbb{E}_\omega\left[\boldsymbol{f}_{i,t} \mid \tilde{I}_{i,t-1}\right] = \boldsymbol{f}_{i,t}$, then

$$\mathbb{E}\left[h_{i,t}\right] = \mathbf{w}_i^\top \boldsymbol{f}_{i,t}, \tag{5}$$

which is a classic multi-factor model at time t.[8]

[7] Jagannathan and Wang (1996) propose that CAPM is valid conditionally, with betas and the market risk premium varying over time, as shown in Equation (2) of their study. Instead of temporal variation, this study focuses on spatial variation in house prices. Thus, the model is adjusted to a spatially conditional CAPM, where the price of risk and risk factors vary across different areas.

[8] Assuming $\boldsymbol{f}_{i,t} = \boldsymbol{f}_t$, meaning risk factors are uniform across different areas, the unconditional expected excess HPI growth rate is:

$$\mathbb{E}\left[h_{i,t}\right] = \mathbf{w}_i^\top \boldsymbol{f}_t, \tag{6}$$

as per the multi-factor CAPM.

Proof. As described by Equation (4) in Jagannathan and Wang (1996), the conditional expectation in Lemma 3.1 can be transformed to the unconditional one in Lemma 3.2. □

The following lemma describes the situation when the complete information set is inaccessible.

Lemma 3.3. *If $I_{i,t-1}(\Omega, \Omega_j, \Omega_{ji})$ is assumed as the full information set, Lemma 3.1 can be rewritten as:*

$$\mathbb{E}\left[h_{i,t}^P \mid I_{i,t-1}(\Omega, \Omega_j, \Omega_{ji})\right] = \mathbb{E}\left[\mathbf{w}_i^P \mid I_{i,t-1}(\Omega, \Omega_j, \Omega_{ji})\right]^\top \mathbb{E}\left[\boldsymbol{f}_{i,t}^P \mid I_{i,t-1}(\Omega, \Omega_j, \Omega_{ji})\right], \tag{7}$$

where $h_{i,t}^P$ is the partial excess HPI growth rate of Suburb i at time t, and the terms $\mathbf{w}_i^P$ and $\boldsymbol{f}_{i,t}^P$ denote the partial risk premia and factors. They are termed "partial" as they are estimated from the partial information set rather than the complete one.

The partial excess HPI growth $\mathbb{E}\left[h_{i,t}^P\right]$ is defined as

$$\begin{aligned}\mathbb{E}\left[h_{i,t}^P\right] &= \mathbb{E}\left[\mathbf{w}_i^P\right]^\top \mathbb{E}\left[\boldsymbol{f}_{i,t}^P\right] \\ &+ \mathbb{C}\mathrm{ov}_\omega\left(\mathbb{E}_\omega\left[\mathbf{w}_i^P \mid I_{i,t-1}(\Omega, \Omega_j, \Omega_{ji})\right], \mathbb{E}\left[\boldsymbol{f}_{i,t}^P \mid I_{i,t-1}(\Omega, \Omega_j, \Omega_{ji})\right]\right).\end{aligned} \tag{8}$$

Here, the covariance term cannot be omitted because the full information set $I_{i,t-1}(\Omega, \Omega_j, \Omega_{ji})$ is incomplete as mentioned in Remark 3.1, leading to randomness in the conditional expectation of risk premia and factors.

According to the structure of data given, the decomposition of the house price model hierarchically by taking conditional expectations of Equation (7) is shown as follows.

Proposition 3.1. *The house price model conditional on the information set $I_{i,t-1}(\Omega_i, \Omega_{ji})$ is written as:*

$$\begin{aligned}&\mathbb{E}\left[h_{i,t}^P \mid I_{i,t-1}\left(\Omega_j, \Omega_{ji}\right)\right] \\ &= \mathbf{w}_i^P([1:\varkappa])^\top \boldsymbol{f}_{i,t}^P([1:\varkappa]) \\ &\quad + \mathbb{E}\left[\mathbf{w}_i^P(-[1:\varkappa]) \mid I_{i,t-1}\left(\Omega_j, \Omega_{ji}\right)\right]^\top \mathbb{E}\left[\boldsymbol{f}_{i,t}^P(-[1:\varkappa]) \mid I_{i,t-1}\left(\Omega_j, \Omega_{ji}\right)\right] \\ &\quad + \mathbb{C}\mathrm{ov}_\omega\left(\mathbb{E}\left[\mathbf{w}_i^P(-[1:\varkappa]) \mid I_{i,t-1}\left(\Omega_j, \Omega_{ji}\right)\right], \mathbb{E}\left[\boldsymbol{f}_{i,t}^P(-[1:\varkappa]) \mid I_{i,t-1}\left(\Omega_j, \Omega_{ji}\right)\right]\right),\end{aligned} \tag{9}$$

where $\mathbf{w}_i^P([1:\varkappa])$ and $\boldsymbol{f}_{i,t}^P([1:\varkappa])$ are the first $\varkappa$ premia and risk factors, and $\mathbf{w}_i^P(-[1:\varkappa])$ and $\boldsymbol{f}_{i,t}^P(-[1:\varkappa])$ denote the premia and factors excluding the first $\varkappa$. Here, $\varkappa$ serves as a threshold for distinguishing risk factors across hierarchical levels.

Proof. See the proof in Appendix A. □

3.2 Hierarchical house price model

We assume that the extracted risk factors, based on partial information, are the same at the same level for different suburbs according to Proposition 3.1. Therefore, risk factors $\boldsymbol{f}_{i,t}^P(\Omega)$ and $\boldsymbol{f}_{i,t}^P(\Omega_j)$ can be written as

$\boldsymbol{f}_t^P(\Omega)$ and $\boldsymbol{f}_t^P(\Omega_j)$. Based on the theoretical analysis above, the expression of the hierarchical house price model is given in this subsection.

Lemma 3.4. *Let $\mathbf{w}_i^P(\Omega)$ and $\mathbf{w}_i^P(\Omega_j)$ denote the vectors of regression coefficients illustrating partial risk premia about national-level and SA4-level risk factors $\boldsymbol{f}_t^P(\Omega)$ and $\boldsymbol{f}_t^P(\Omega_j)$ in the previous periods for Suburb i. A two-level hierarchical model is presented below, which captures the relationship between the house prices and exogenous variables.*

$$\begin{aligned}
&\textit{HPI of suburb } i \in \Omega_j : h_{i,t}^P(\Omega_j) = l_t + l_{j,t} + \boldsymbol{\beta}_i^P(\Omega)^\top \boldsymbol{f}_t^P(\Omega) \sum_{\kappa>0}^{K} L^\kappa + \boldsymbol{\beta}_i^P(\Omega_j)^\top \boldsymbol{f}_t^P(\Omega_j) \sum_{\kappa>0}^{K} L^\kappa, \\
&\textit{Average HPI} : l_t = \bar{\mathbf{w}}_i^P(\Omega)^\top \boldsymbol{f}_t^P(\Omega) \sum_{\kappa>0}^{K} L^\kappa + \eta_t, \\
&D(\Omega, \Omega_j) : l_{j,t} = \bar{\mathbf{w}}_i^P(\Omega_j)^\top \boldsymbol{f}_t^P(\Omega_j) \sum_{\kappa>0}^{K} L^\kappa + \eta_{j,t}.
\end{aligned} \tag{10}$$

Here $\bar{\mathbf{w}}_i^P(\Omega)$ is the average of $\mathbf{w}_i^P(\Omega)$ and $\boldsymbol{\beta}_i^P(\Omega)$ is the deviation of $\mathbf{w}_i^P(\Omega)$ from the average, that is, $\mathbf{w}_i^P(\Omega) - \bar{\mathbf{w}}_i^P(\Omega)$; $\bar{\mathbf{w}}_i^P(\Omega_j)$ and $\boldsymbol{\beta}_i^P(\Omega_j)$ are similarly defined; L represents the lag operator, with the lag order denoted by a positive integer $\kappa \leq K$; $D(\cdot,\cdot)$ measures the distance between HPI growths at different levels; η_t and $\eta_{j,t}$ represent random residuals with finite mean and variance.

Proof. See the proof in Appendix A. □

Lemma 3.4 verifies the basic idea of this hierarchical model: collecting information at smaller scales to analyse residuals. Therefore, a three-level model can be accomplished if more information is collected at the most granular scale. However, the estimation error for numerous risk factors could be significant due to the limited data available for each postcode-level area. Therefore, only the national and SA4 levels are considered for building a two-level hierarchical model before analysing residuals.[9]

Remark 3.2. *The relationship between the real house price growth index and the prediction from the two-level model is:*

$$h_{i,t} = h_{i,t}^P(\Omega_j) + e_{i,t}(\Omega_j) + \psi_t + \psi_{j,t}, \tag{11}$$

where ψ_t and $\psi_{j,t}$ are the error terms arising from the formulation of η_t and $\eta_{j,t}$. The residuals independent of the distribution assumption of η_t and $\eta_{j,t}$ are denoted as $e_{i,t}(\Omega_j)$.

Proof. See the proof in Appendix A. □

Proposition 3.2 further discusses the residual $e_{i,t}(\Omega_j)$ of the two-level hierarchical house price model for the excess HPI growth rate of the i^{th} suburb, highlighting the errors that the model cannot explain at granular scales due to data collection limitations.

[9] Another interpretation is that the two-level hierarchical model is equivalent to a three-level model with $D(\Omega_j, \Omega_{ji}) = 0$ due to the unavailability of obtaining $\boldsymbol{f}_{i,t}^P(\Omega_{ji})$ as an estimate of the corresponding risk factors in area i. The definition of $D(\cdot,\cdot)$ is provided in Lemma 3.4, where $D(\Omega_j, \Omega_{ji})$ represents the difference in excess HPI growth between the SA4 and postcode levels.

Proposition 3.2. *The lack of information will lead to inaccurate estimates of the two-level hierarchical house price models. The difference between the mean value of house prices and forecasts are*

$$\begin{aligned} &\mathbb{E}\left[h_{i,t}\right]-\mathbb{E}_{\omega}\left[h_{i,t}^{P}\right] \\ =&\,\mathbb{E}\left[\mathbf{w}_{i}\right]^{\top}\mathbb{E}\left[\boldsymbol{f}_{i,t}\right]\sum_{\kappa>0}L^{\kappa}-\mathbb{E}[\mathbf{w}_{i}^{P}]^{\top}\mathbb{E}\left[\boldsymbol{f}_{i,t}^{P}\right]\sum_{\kappa>0}L^{\kappa} \\ &-\mathbb{C}\mathrm{ov}_{\omega}\left(\mathbb{E}\left[\mathbf{w}_{i}^{P}\mid I_{i,t-1}(\Omega,\Omega_{j},\Omega_{ji})\right],\mathbb{E}\left[\boldsymbol{f}_{i,t}^{P}\sum_{\kappa>0}L^{\kappa}\mid I_{i,t-1}(\Omega,\Omega_{j},\Omega_{ji})\right]\right), \end{aligned} \tag{12}$$

and the final error of the hierarchical model $e_{i,t}(\Omega_j)$ can be expressed as

$$e_{i,t}(\Omega_j)=\underbrace{\mathbb{E}\left[h_{i,t}\right]-\mathbb{E}_{\omega}[h_{i,t}^{P}]}_{\textit{Time-variant trend term}}-\underbrace{(\eta_t+\eta_{j,t}+\psi_t+\psi_{j,t}-\tilde{\epsilon}_{i,t})}_{\textit{Combination of random noises}}, \tag{13}$$

where $\tilde{\epsilon}_{i,t}$ represents unspecified random noise with a mean of 0 and finite variance, capturing the deviation between the expected house price growth index predicted by the true model and the observed real value.

Proof. See the proof in Appendix A. □

Equation (13) in Proposition 3.2 indicates that the final errors of the hierarchical model in each area are time series, and the linear stochastic difference only consists of two distinct parts: a time-variant trend term and a combination of random noises. Based on this result, the final hierarchical model is derived and shown in Proposition 3.3.

Proposition 3.3. *The final three-level hierarchical model can be written as follows:*

$$\begin{aligned} &\textit{HPI of suburb } i\in\Omega_j: h_{i,t}^{P}(\Omega_j)=l_t+l_{j,t}+l_{ji,t}, \\ &\textit{Average HPI}: l_t=\bar{\mathbf{w}}_{i}^{P}(\Omega)^{\top}\boldsymbol{f}_{t}^{P}(\Omega)\sum_{\kappa>0}L^{\kappa}+\eta_t, \\ &D\left(\Omega,\Omega_j\right): l_{j,t}=\bar{\mathbf{w}}_{i}^{P}(\Omega_j)^{\top}\boldsymbol{f}_{t}^{P}(\Omega_j)\sum_{\kappa>0}L^{\kappa}+\eta_{j,t}, \\ &D\left(\Omega_j,\Omega_{ji}\right): l_{ji,t}=\boldsymbol{\beta}_{i}^{P}(0)+\boldsymbol{\beta}_{i}^{P}(\Omega)^{\top}\boldsymbol{f}_{t}^{P}(\Omega)\sum_{\kappa>0}L^{\kappa}+\boldsymbol{\beta}_{i}^{P}(\Omega_j)^{\top}\boldsymbol{f}_{t}^{P}(\Omega_j)\sum_{\kappa>0}L^{\kappa}. \end{aligned} \tag{14}$$

The final observation equation[10] *of the hierarchical model is*

$$h_{i,t}^{P}(\Omega_j)=\boldsymbol{\beta}_{i}^{P}(0)+l_t+l_{j,t}+\boldsymbol{\beta}_{i}^{P}(\Omega)^{\top}\boldsymbol{f}_{t}^{P}(\Omega)+\boldsymbol{\beta}_{i}^{P}(\Omega_j)^{\top}\boldsymbol{f}_{t}^{P}(\Omega_j). \tag{15}$$

Proof. See the proof in Appendix A. □

[10]The observation equation only contains the current observations of variables.

3.3 Risk factors and joint dynamics

We assume that the risk factors generated from exogenous variables, along with l_t ($l_{j,t}$), as defined in Proposition 3.3, follow VAR models:

$$\begin{aligned} \boldsymbol{F}_t^P(\Omega) &= \Psi_L(\Omega)\boldsymbol{F}_t^P(\Omega) + H_t(\Omega), \\ \boldsymbol{F}_t^P(\Omega_j) &= \Psi_L(\Omega_\mathrm{j})\boldsymbol{F}_t^P(\Omega_j) + H_t(\Omega_j), \end{aligned} \tag{16}$$

where L is the lag operator, $\Psi_L(\Omega)$ is the lag polynomial matrix at the national level, and $\Psi_L(\Omega_j)$ is the lag polynomial matrix for area j at the SA4 level. $\boldsymbol{F}_t^P(\Omega)$ (resp. $\boldsymbol{F}_t^P(\Omega_j)$) is a vector of extended risk factors, consisting of the original risk factor vector $\boldsymbol{f}_t^P(\Omega)$ (resp. $\boldsymbol{f}_t^P(\Omega_j)$) and l_t (resp. $l_{j,t}$). Similarly, H_t (resp. $H_{j,t}$) is an extended vector of errors for all risk factors, including η_t (resp. $\eta_{j,t}$).

This hierarchical multi-factor model explores optimal strategies for selecting risk factors. PCA is employed to estimate risk factors at different levels, enabling the incorporation of more exogenous information to analyse prediction residuals from the upper-level model.

3.3.1 Risk factors at the national level

We determine the optimal parameters for the national-level VAR model using three standard information criteria: Akaike's information criterion (AIC), Schwarz–Bayesian information criterion (BIC), and Hannan–Quinn information criterion (HQIC). For parsimony, VAR(1) is selected as the preferred model at the national level, with a comparison of various models provided in Table 9 in Appendix B. Different orders of vector autoregressive moving average (VARMA) models are also evaluated in that table, but these VARMA models do not outperform the VAR models across the criteria. The estimated parameters and the Cholesky decomposition of the error covariance matrix for the VAR(1) model are likewise presented in Table 10 in Appendix B. Appendix C includes the PCA loading matrix at the national level and discusses the suitability of using PCA in this context.

3.3.2 Risk factors at the SA4 level

To account for potential variations in the construction of risk factors, Lettau and Pelger (2020) propose the RP-PCA method for estimating factors in high-dimensional data using statistical factor analysis. RP-PCA is a modified version of PCA with a penalty term for pricing errors. The RP-PCA model is applied to each SA4 area, Ω_j, to extract risk factors from the collected variables $X_t(\Omega_j)$.

The adapted RP-PCA penalty function includes two terms that capture the overall cross-time and cross-area errors of the PCA method:

$$(\widehat{\boldsymbol{f}}^P(\Omega_j), \widehat{\Lambda}) = \underset{\boldsymbol{f}^P(\Omega_j),\Lambda}{\text{argmin}} \left\{ \underbrace{\frac{1}{n(\Omega_j \in \Omega)T} \sum_{\Omega_j \in \Omega} \sum_{t=1}^{T} |X_t(\Omega_j) - \boldsymbol{f}_t^P(\Omega_j)\Lambda^\top|^2}_{\text{unexplained variation}} \right.$$
$$\left. + \underbrace{\gamma_1 \frac{1}{T} \sum_{t=1}^{T} |\overline{X}_t - \overline{\boldsymbol{f}}_t^P \Lambda^\top|^2}_{\text{cross-area error}} + \underbrace{\gamma_2 \frac{1}{n(\Omega_j \in \Omega)} \sum_{\Omega_j \in \Omega} |\overline{X}(\Omega_j) - \overline{\boldsymbol{f}}^P(\Omega_j)\Lambda^\top|^2}_{\text{cross-time error}} \right\}.$$

The two penalty terms in the RP-PCA method are denoted by γ_1 and γ_2. $\overline{X}(\Omega_j)$ and $\overline{\boldsymbol{f}}_i^P(\Omega_j)$ correspond to the average variables and risk factors across time in area j. In practical applications of the PCA method, the data matrix, X, is demeaned by subtracting the sample mean from each observation across time. As a result, the second penalty term γ_2 is set to be zero, and the objective function can be simplified to Equation (17). When the weight of the second penalty term γ_2 is set to be zero, the tuning parameter γ_1 becomes the only parameter that needs to be adjusted.

The estimated risk factors $\widehat{\boldsymbol{f}}^P(\Omega_j)$ and the corresponding loading matrix $\widehat{\Lambda}$ are obtained according to the following rule:

$$(\widehat{\boldsymbol{f}}^P(\Omega_j), \widehat{\Lambda}) = \underset{\boldsymbol{f}^P(\Omega_j),\Lambda}{\text{argmin}} \left\{ \underbrace{\frac{1}{n(\Omega_j \in \Omega)T} \sum_{\Omega_j \in \Omega} \sum_{t=1}^{T} |X_t(\Omega_j) - \boldsymbol{f}_t^P(\Omega_j)\Lambda^\top|^2}_{\text{unexplained variation}} + \underbrace{\gamma_1 \frac{1}{T} \sum_{t=1}^{T} |\overline{X}_t - \overline{\boldsymbol{f}}_t^P \Lambda^\top|^2}_{\text{cross-area error}} \right\}. \quad (17)$$

The penalty term in the RP-PCA method is denoted by γ_1, while Λ represents the PCA loadings for the middle-level model. The terms $\overline{X}_t$ and $\overline{\boldsymbol{f}}_t^P$ refer to the average variables and risk factors across areas at time t, respectively. The parameter γ_1 is tunned to either underweighting or overweighting the means across different areas.

1. As the tuning parameter γ_1 approaches infinity, the means across different areas are overweighted, and the objective function given in Equation (17) is simplified to:

$$(\widehat{\boldsymbol{f}}^P(\Omega_j), \widehat{\Lambda}) = \underset{\Lambda,\boldsymbol{f}^{\boldsymbol{P}}(\boldsymbol{\Omega_j})}{\text{argmin}} \left\{ \frac{1}{T} \sum_{t=1}^{T} |\overline{X}_t - \overline{\boldsymbol{f}}_t^P \Lambda^\top|^2 \right\}. \quad (18)$$

Minimising Equation (18) is equivalent to applying PCA to the covariance matrix of the centralised average variables $\bar{X}_t$ across time, provided that the average variables across different areas have been demeaned. The motivation for overweighting the means across different areas at each time point is to reduce the amount of risk factors.

2. When the tuning parameter γ_1 is set to be zero, the objective function given in Equation (17) is:

$$(\widehat{\boldsymbol{f}}^P(\Omega_j), \widehat{\Lambda}) = \underset{\Lambda, \boldsymbol{f}^P(\boldsymbol{\Omega}_j)}{\operatorname{argmin}} \left\{ \frac{1}{n(\Omega_j \in \Omega)T} \sum_{\Omega_j \in \Omega} \sum_{t=1}^{T} |X_t(\Omega_j) - \boldsymbol{f}_t^P(\Omega_j)\Lambda^\top|^2 \right\}. \tag{19}$$

In this case, the means across different areas are underweighted, indicating that the across-area errors are no longer considered when identifying the principal components. Consequently, the dimension of loadings will expand, leading to a substantial rise in the number of risk factors.

3. When the tuning parameter γ_1 is set to be -1, the objective function given in Equation (17) can be rewritten as:

$$\begin{aligned}&(\widehat{\boldsymbol{f}}^P(\Omega_j), \widehat{\Lambda}) \\ &= \underset{\Lambda, \boldsymbol{f}^P(\Omega_j)}{\operatorname{argmin}} \left\{ \frac{1}{n(\Omega_j \in \Omega)T} \sum_{\Omega_j \in \Omega} \sum_{t=1}^{T} \left(\left(X_t(\Omega_j) - \overline{X}_t \right) - \left(\boldsymbol{f}_t^P(\Omega_j) - \overline{\boldsymbol{f}}_t^P \right) \Lambda^\top \right)^2 \right\}.\end{aligned} \tag{20}$$

In this case, the cross-area means of the variables and risk factors are subtracted from the original data to partially eliminate the data trend. This approach is commonly used to handle non-stationary time series. However, the tuning parameter γ_1 is considered greater than zero in the subsequent analysis for two main reasons. Firstly, as previously mentioned in Section 2, all variables have been pre-processed to ensure stationarity before the analysis. Moreover, assigning positive values to γ_1 facilitates easier interpretation.

The final objective function is defined as[11]:

$$\underset{\Lambda, \boldsymbol{f}^P(\boldsymbol{\Omega}_j)}{\operatorname{argmin}} \left\{ \underbrace{\frac{1}{n(\Omega_j \in \Omega)T} \sum_{\Omega_j \in \Omega} \sum_{t=1}^{T} |X_t(\Omega_j) - \boldsymbol{f}_t^P(\Omega_j)\Lambda^\top|^2}_{\text{unexplained variation}} + \underbrace{\frac{\gamma(\gamma+2)}{T} \sum_{t=1}^{T} |\overline{X}_t - \overline{\boldsymbol{f}}_t^P \Lambda^\top|^2}_{\text{cross-area error}} \right\}. \tag{21}$$

When γ is greater than or equal to zero, the product $\gamma(\gamma+2)$ increases monotonically from zero to infinity. After assuming $\gamma_1 = \gamma(\gamma+2)$, adjusting the tuning parameter γ in the range [0,+∞] for Equation (21) is essentially equivalent to adjusting the tuning parameter γ_1 in the range [0,+∞] for Equation (17). An increase in γ signifies a greater emphasis on the importance of cross-area error.

The validation method we employ tunes the penalty term, denoted by γ, over the range $[0, +\infty]$. Let $d_{90\%}$ denote the minimum number of risk factors that explain at least 90% of the variance. For each γ

[11]Equation (21) can be rewritten as

$$(\widehat{\boldsymbol{f}}^P(\Omega_j), \widehat{\Lambda}) = \underset{\Lambda, \boldsymbol{f}^P(\Omega_j)}{\operatorname{argmin}} \left\{ \frac{1}{n(\Omega_j \in \Omega)T} \sum_{\Omega_j \in \Omega} \sum_{t=1}^{T} \left(\left(X_t(\Omega_j) + \gamma\overline{X}_t \right) - \left(\boldsymbol{f}_t^P(\Omega_j) + \gamma\overline{\boldsymbol{f}}_t^P \right) \Lambda^\top \right)^2 \right\},$$

which is equivalent to Equation (17) when the tuning parameter γ (γ_1) is set to be infinity, 0, and -1, as shown in Equations (18), (19), and (20).

value, it identifies the number of risk factors $d_{90\%}$ and establishes a consistent method for constructing risk factors across all SA4 structures. As γ increases, $d_{90\%}$ tends to decrease and eventually stabilises. This relationship, including the implications for the number of principal components and the application of the AIC in selecting the optimal number of components, is intricately connected to the limitations of the VAR model and the data's characteristics. The optimal selection of risk factors for each γ value is determined through recursive feature elimination (RFE), considering various constraints. All details, including technical nuances and step-by-step procedures, are thoroughly explained in Appendix B.

3.4 Residual analysis in each suburb

This subsection provides a detailed analysis of the residuals of the expected HPI growths of the actual model $\tilde{\epsilon}_{i,t}$, defined in Proposition 3.2. As $\boldsymbol{\beta}_i^P(0)$ has been proved as time-invariant and integrated into the final model presented in Proposition 3.3, Equation (13) is rewritten as:

$$\tilde{\epsilon}_{i,t} = e_{i,t}(\Omega_j) + \psi_t + \psi_{j,t} + \epsilon_t + \epsilon_{j,t}, \tag{22}$$

which measures the difference between the actual and predicted HPI growth rates. Despite the stability of the distributional properties of observed residuals, their distribution in different areas may be diverse and not normally distributed. Under the stability assumption, multi-dimensional empirical copulas will generate pseudo-random samples, modelling the dependence structure between the residuals in different suburbs. These simulated results forecast future house price growth rates at different geographical levels (Ho et al., 2016).

Figure 2 depicts the flow chart of the simulation process for forecasting with residual analysis. The simulation process involves obtaining various levels of risk factors based on available information, using corresponding VAR models for the period in question. An unbiased estimate is calculated as the sum of the predicted average HPI growth rates for the national and SA4 levels, in addition to an expected value derived from the linear model of the obtained risk factors. When residual analysis is incorporated into the model, the unbiased estimate is augmented with stochastic errors generated based on multivariate copulas across different suburbs. The simulation outputs the final results of the forecast in one period, including errors in predicted average HPI growth rates at the national and SA4 levels. The values of l and l_j are then obtained by averaging all the simulated results, and the corresponding error terms are computed. The detailed simulation process is presented in Algorithm 1.

4 Numerical results

This section evaluates the model developed in earlier sections. First, the robustness of the RP-PCA method is assessed, followed by an exploration of how residual analysis affects model performance. The model is then

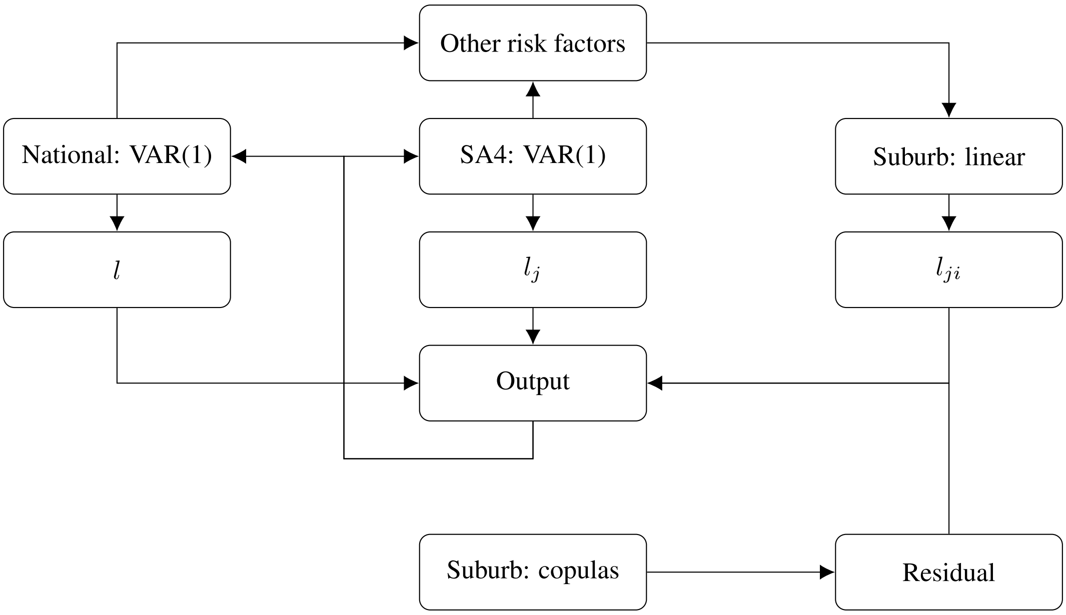


Figure 2. Flow chart of the forecasting simulation process with residual analysis. This diagram details the steps from obtaining risk factors and using VAR models to generating an unbiased estimate of average HPI growth rates at the national and SA4 levels. It also illustrates the incorporation of stochastic errors through multivariate copulas for residual analysis.

Algorithm 1 Simulation algorithm for the hierarchical model with empirical copulas.

Require: $T > 0$ ▷ T denotes the terminal time point for the forecasts
  $s \leftarrow 1$ ▷ s represents the index of simulations
  while $s \leq S$ **do** ▷ S represents the total number of simulations
    Fit VAR models for l and l_j and the risk factors for every j before time t.
    while $t \leq T$ **do** ▷ t denotes the beginning time point for the forecasts
      Estimate $\mathbb{E}[l_{t+1}]$ and $\mathbb{E}[l_{j,t+1}]$ based on VAR models.
      Estimate $\mathbb{E}[h^P_{i,t+1}(\Omega_j)]$ according to Equation (30) for every i and j.
      Generate random errors based on the one-step-ahead error using empirical copulas
      Obtain $h^P_{i,t+1}(\Omega_j)$ by adding the one of randomly generated one-step-ahead errors
      Obtain $l_{0,t+1}$ and $l_{j,t+1}$ based on $h^P_{i,t+1}(\Omega_j)$
      Update the corresponding VAR model by incorporating the $l_{0,t+1}$ and $l_{j,t+1}$
      $t = t + 1$
    end while
  end while
  $s = s + 1$

---

simplified by excluding external variables to emphasise its core mechanisms. A comparison of the simplified model's predictive accuracy with the MinT(Shrink) model (Wickramasuriya et al., 2018) highlights relative strengths and weaknesses. We also compare the hierarchical and FAVAR models Bernanke et al. (2005), focusing on their interpretive and predictive capabilities for HPI data. The discussion also examines the advantages and economic relevance of the RP-PCA model. Further analysis investigates risk factors at national and SA4 levels, highlighting their characteristics and associated risks. Finally, a sensitivity analysis provides economic insights.

4.1 Robustness of the adapted RP-PCA

In the RP-PCA model, the loadings of risk factors provide insights into the contribution of variables to each component. To test the model's robustness, we add five randomly generated time series as disturbances to the variables in each SA4 area to the original observable data $X_t(\Omega_j)$ in Equation (17).[12] Table 3 presents the resulting average loadings. By comparing the relative changes in the loadings of variables and the new loadings of the disturbances, we can validate the robustness of the RP-PCA method.

4.2 Results of the residual analysis of the full hierarchical model

We first justify the integration of empirical copulas into the complete hierarchical model. While the residuals of the one-step-ahead forecasts from the full hierarchical model exhibit stable distributional properties overall, their distributions differ across suburbs and deviate from normality. The Kolmogorov-Smirnov and the Shapiro-Wilk tests reject the null hypothesis that residuals follow a normal distribution at the 1% significance level across all suburbs.

Figure 3 illustrates the residual distributions in six randomly selected suburbs, arranged in ascending order of skewness from top to bottom. These distributions deviate notably from normality, consistent with the statistical test results. Additionally, a positive correlation between skewness and kurtosis is evident.

The correlation heatmap of residuals across all suburbs is presented in Figure 4, with grids arranged by the postcodes of the suburbs. Postcode similarity is assumed to approximate geographical proximity. A strong correlation is generally observed among suburbs with more similar postcodes. Notably, positive correlation coefficients are primarily found in adjacent suburbs, illustrated by small knots along the diagonal in Figure 4. In contrast, residual correlations in less adjacent suburbs are negative, as depicted by green squares symmetrically distributed around the diagonal. Correlations between distant areas tend to approach zero.

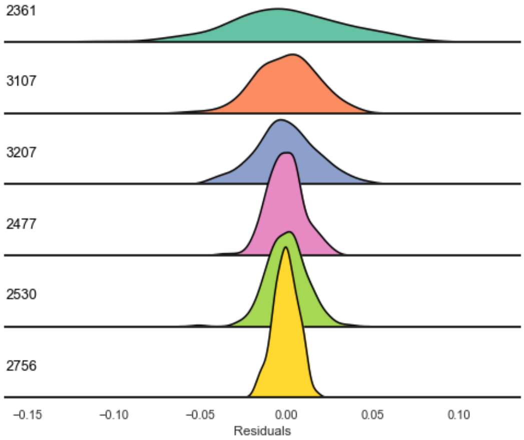
Next, we present the outcomes of integrating the empirical copula. The probability density function of average HPIs is simulated. Figure 5 displays the probability density functions of the simulated data with and without copulas alongside the density function of the actual historical data. The comparison reveals

[12]The number of disturbances is set to 20% of the original variables.

Table 3. Robustness test: the average loadings of risk factors at the SA4 level.

Variable	Risk factor 1	Risk factor 2	Risk factor 3	Risk factor 4	Risk factor 5
hsdebt	-0.049001	-0.054618	0.084474	-0.373761	0.079578
	(-0.72%)	(-2.65%)	(-2.62%)	(-1.32%)	(1.17%)
wsce	0.040468	0.044586	0.189539	0.053279	-0.193295
	(-3.26%)	(-1.36%)	(2.07%)	(0.49%)	(-2.12%)
agemedian	-0.288531	-0.049618	0.060317	-0.164907	-0.277348
	(-3.81%)	(-0.27%)	(-3.78%)	(-1.18%)	(1.98%)
agestd	0.421709	-0.014769	0.241167	0.014759	0.105119
	(-1.07%)	(-3.05%)	(0.18%)	(-0.98%)	(0.23%)
ageq3	0.127716	0.068639	-0.161414	-0.273830	0.424677
	(-5.96%)	(-1.86%)	(-1.06%)	(3.82%)	(-0.38%)
agecount	-0.153737	-0.010077	-0.013228	-0.079194	-0.095319
	(-0.38%)	(-0.72%)	(-1.31%)	(0.08%)	(0.44%)
xphmrna	0.117442	-0.117297	0.498115	-0.057740	-0.149872
	(-1.38%)	(-0.57%)	(-0.10%)	(1.33%)	(1.89%)
xphltpa	0.251271	-0.645910	-0.265797	-0.371495	-0.080600
	(0.12%)	(-1.89%)	(0.57%)	(0.40%)	(-0.17%)
lsemp	-0.222298	-0.498821	0.046789	0.372046	0.414242
	(0.31%)	(0.40%)	(0.22%)	(0.35%)	(1.40%)
iprbwm	0.133836	0.355005	0.077661	0.067055	0.055210
	(-0.79%)	(-0.51%)	(0.04%)	(1.63%)	(0.08%)
fiprbmr	0.106791	0.034247	-0.069302	-0.021411	0.048765
	(-0.40%)	(-2.30%)	(-1.84%)	(-1.07%)	(0.26%)
ancob_1101	-0.083466	0.100582	-0.158472	0.073831	0.072029
	(0.29%)	(0.69%)	(0.45%)	(0.06%)	(1.06%)
ancob_1201	0.148425	0.090570	0.133342	0.167515	-0.025991
	(1.06%)	(-0.85%)	(1.15%)	(1.78%)	(0.33%)
ancob_7103	-0.208439	0.169382	0.108600	-0.093309	-0.005138
	(-1.31%)	(0.22%)	(1.37%)	(0.76%)	(-0.35%)
ancob_6101	-0.264256	-0.039849	0.001015	-0.186877	-0.073736
	(-1.64%)	(1.32%)	(-0.94%)	(-1.38%)	(-0.80%)
ancob_5105	0.205965	-0.044643	-0.115827	0.143242	0.148128
	(-0.95%)	(0.50%)	(-0.51%)	(0.60%)	(0.91%)
edhigh1_4	-0.067241	0.212654	-0.621667	0.099669	-0.104623
	(-0.56%)	(0.97%)	(-0.22%)	(1.99%)	(-0.13%)
edhigh1_5	-0.130216	-0.073308	-0.032706	-0.099762	-0.315203
	(-0.31%)	(1.35%)	(-0.61%)	(-0.65%)	(1.56%)
edhigh1_8	-0.019074	-0.092077	-0.127864	0.060256	0.055505
	(-1.16%)	(1.25%)	(1.46%)	(-1.39%)	(0.84%)
edhigh1_9	-0.179356	-0.117898	0.156384	-0.075394	-0.114189
	(1.02%)	(0.58%)	(1.47%)	(-0.30%)	(0.71%)
anatsi_2	-0.269164	-0.184877	0.071526	0.481738	0.019514
	(1.69%)	(-1.03%)	(-0.93%)	(-0.33%)	(0.84%)
anatsi_3	-0.479766	0.100232	0.078930	-0.247900	0.344730
	(-0.03%)	(0.74%)	(0.23%)	(0.09%)	(1.27%)
chkb12_1	0.004765	-0.105922	-0.179660	0.203384	-0.444280
	(2.95%)	(-0.77%)	(2.36%)	(0.87%)	(1.11%)
Disturbance_1	0.000536	-0.000117	0.002385	-0.000285	0.001253
Disturbance_2	0.001308	0.001256	0.002847	-0.002580	0.001410
Disturbance_3	-0.004404	-0.000404	0.001973	0.003849	0.000927
Disturbance_4	-0.001619	0.004423	0.000447	0.001142	0.003052
Disturbance_5	0.003230	-0.002180	-0.000744	0.003263	0.000414

Notes: Above the dashed line are loadings after introducing disturbances and the relative changes compared to the previous loadings of variables introduced in Table 2. Below the dashed line are the average loadings of the newly introduced disturbances, which are randomly generated.



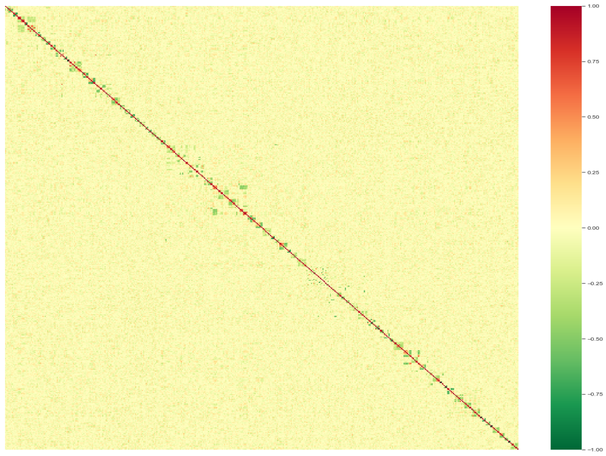


Figure 3. Distributions of residuals in six selected suburbs. This figure illustrates the residual distributions for the suburbs of Ashford, Templestowe Lower, Port Melbourne, Alstonville, Dapto, and Windsor, arranged in ascending order of skewness from top to bottom. The corresponding postcodes are displayed.

Figure 4. Correlation heatmap of residuals across suburbs. The heatmap illustrates positive correlations for adjacent suburbs (tiny knots on the diagonal) and negative correlations for less adjacent ones (green squares around the diagonal). Correlations approach zero for more distant suburbs, indicating a geographical dependency.

Ashford is located in South Australia, Templestowe Lower and Port Melbourne in Victoria, Alstonville and Dapto in New South Wales, and Windsor in Queensland.

that the slight deviation in the simulated results is corrected upon incorporating the residual analysis.

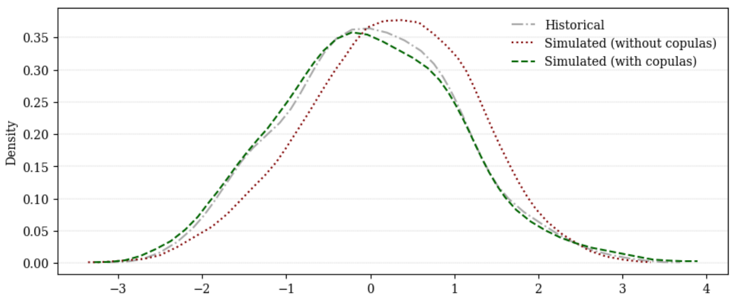


Figure 5. Impact of empirical copulas on HPI simulation. This figure contrasts the probability density functions of average HPIs–simulated with and without empirical copulas–against the density function of real historical data. The side-by-side comparison addresses the slight deviations in simulated results, illustrating the correction implied by including residual analysis.

Following this, we conduct simulations to estimate the expected shortfalls of HPI at the postcode level across various confidence levels. Figure 6 illustrates the relative changes in expected shortfalls simulated with and without incorporating copulas after five years (December 2025), based on HPI data from December 2020. The figure contains three graphs in Category (a), which depict results with residual analysis at different confidence levels, and three graphs in Category (b), which present results without residual analysis. Category (b) consistently yields higher house price estimates in low house price scenarios. Additionally, the predicted shortfalls in Category (a) exhibit greater fluctuations across different areas compared to those in Category (b). This indicates that models incorporating residual analysis tend to underestimate HPI in worst-case scenarios, producing more conservative predictions. These findings align with the results presented in Figure 5.

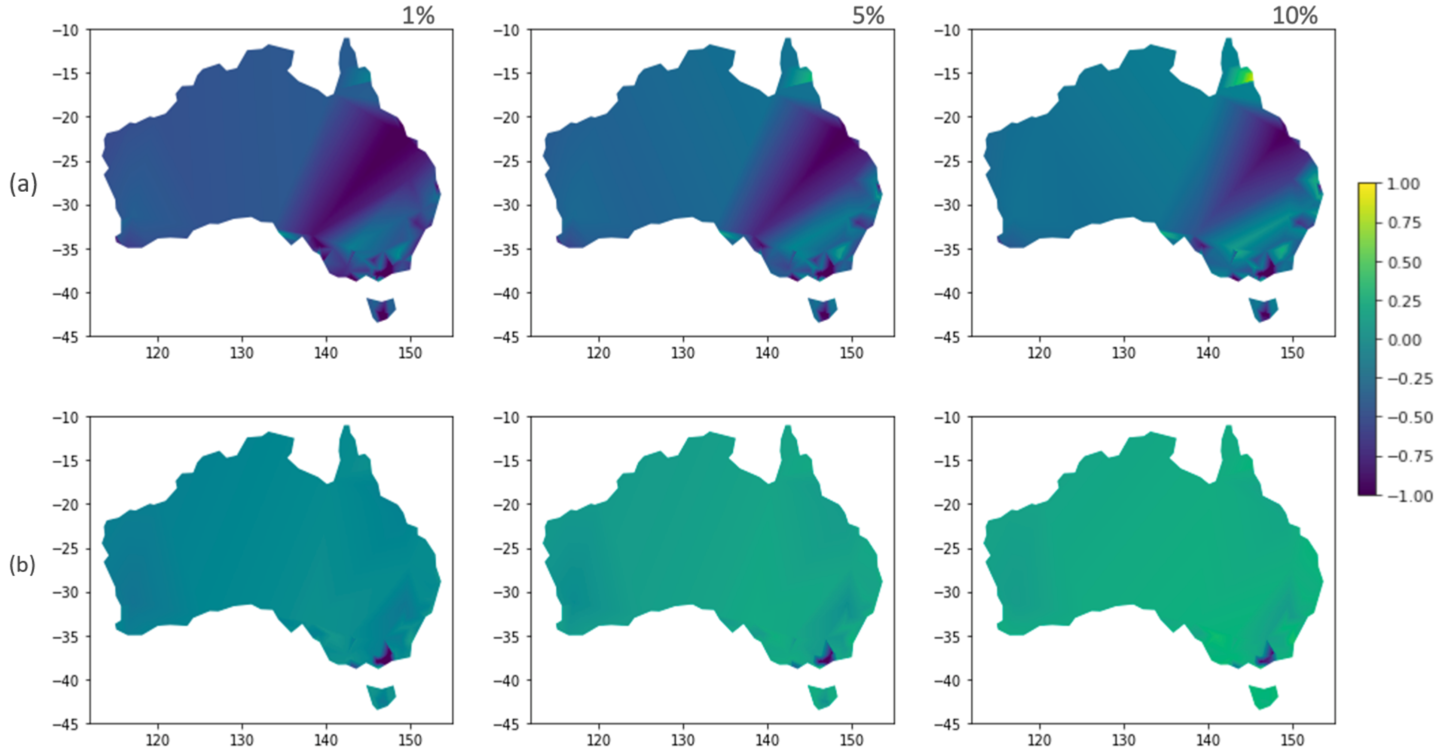


Figure 6. Comparative analysis of expected shortfalls in HPI across different suburbs at 99%/95%/90% confidence levels. This figure illustrates the relative changes in expected HPI shortfalls at the postcode level, simulated over a 5-year horizon from December 2020 with and without the incorporation of residual analysis. Category (a) shows the fluctuations at three confidence levels with residual analysis, while Category (b) depicts outcomes without residual analysis. This indicates that models incorporating residual analysis tend to produce more conservative predictions.

In conclusion, multivariate copulas are employed because the residuals from the hierarchical model exhibit interdependencies across suburbs, and their marginal distributions deviate from normality. By integrating multiple copulas, prediction errors across postcode-level areas are mitigated. Furthermore, residual analysis contributes to more conservative long-term HPI forecasts.

4.3 Comparison of the hierarchical model with the MinT(Shrink) model

The hierarchical model for predicting house prices $h^P_{i,t}(\Omega_j)$ can be simplified by excluding the influence of exogenous variables, resulting in a concise form where l_t and $l_{j,t}$ are estimated using autoregressive integrated moving average (ARIMA) models. In each simulation, the final predicted house prices are generated by adding the estimated value of $h^P_{i,t}(\Omega_j)$ to a marginal copula produced from the covariance of the one-step-ahead errors.

To evaluate the effectiveness of the simplified hierarchical model and its residual analysis, a comparison is made with the MinT(Shrink) model proposed in Wickramasuriya et al. (2018). The MinT(Shrink) model employs a hierarchical structure for forecasting time series data, similar to our approach. It first segments the time series data into clusters, fits individual time series models to each cluster, and then combines these models hierarchically to produce forecasts for the overall series. Both models also account for the impact of one-step-ahead errors in the base forecasts.

The rolling window approach compares the predictive accuracy of the simplified hierarchical model and the MinT(Shrink) model. The performance of four models—the MinT(Shrink) model, the base model of MinT(Shrink), the hierarchical model with empirical copulas, and the hierarchical model without empirical copulas—is evaluated over time using the root mean squared error (RMSE).[13]

RMSE for one to twelve steps ahead forecasts are generated using a training window of 120 observations. The base forecasts of the MinT(Shrink) model and the hierarchical model without empirical copulas are produced through ARIMA models, fitted under default settings in the automated algorithms described by Hyndman and Khandakar (2008) and implemented in the `statsmodels` package for Python. The outcomes of both the base and MinT(Shrink) models are validated by employing the `hts` package in R (Wang, 2021).

Table 4. Comparative table of RMSE for one to twelve steps ahead forecasts using predictive models.

	1	2	3	4	5	6
Base($\times 10^{-2}$)	2.29	2.41	2.47	2.46	2.48	2.49
MinT(S)(%)	-10.89	-9.26	-6.34	-4.38	-2.35	-0.62
Hier(%)	-3.74	-2.64	-0.70	0.34	1.60	2.83
Hier(C)(%)	-10.34	-10.44	-9.19	-7.56	-6.22	-4.87
	7	**8**	**9**	**10**	**11**	**12**
Base($\times 10^{-2}$)	2.48	2.47	2.46	2.47	2.49	2.52
MinT(S)(%)	0.76	1.59	1.68	1.40	0.79	0.08
Hier(%)	3.74	4.16	3.67	2.89	1.76	0.90
Hier(C)(%)	-3.58	-3.20	-3.74	-4.91	-5.29	-6.05

Notes: The table presents RMSE values of four models: the base model of MinT(Shrink), the MinT(Shrink) model, the hierarchical model excluding empirical copulas, and concluding with the hierarchical model incorporating empirical copulas. The 'Base' line indicates the average RMSE (multiplied by 10^{-2}) of the baseline forecasts. Values below the 'Base' line represent the relative percentage decrease (negative values) or increase (positive values) in average RMSE compared to the forecasts of the base model.

As observed from Table 4, the MinT(Shrink) model outperforms other models when the forecasting horizon is one step ahead. However, its relative RMSE compared to the base model transitions from negative to positive as the forecast horizon increases, indicating a decline in performance. This could be due to the model's inability to capture the complex dependencies among the residuals under the Gaussian assumption over longer periods.

[13]The MinT(Shrink) model is built upon the base model, whereas the hierarchical model (without residual analysis) serves as the basis for the hierarchical model (with residual analysis). In the univariate case, the base model characterises the dependent variable's average, whereas the hierarchical model emphasises the distance of this average across various levels. At the lowest level, the hierarchical model uses a linear model to incorporate exogenously collected risk factors, which are constant in the univariate case. In contrast, the base model continues to rely on ARIMA models to describe HPIs. Another reason the hierarchical model (with residual analysis) differs from the MinT(Shrink) model is that it employs a numerical method integrating the one-step-ahead error to construct a numerical model. However, this numerical approach in the hierarchical model carries a limitation. Unlike the MinT(Shrink) model, which theoretically guarantees the minimum variance of the estimation under the assumption that errors conform to a multivariate Gaussian distribution, the empirical copulas employed in the hierarchical model cannot offer such theoretical guarantees with its numerical approach.

Conversely, the hierarchical model with empirical copulas performs best for forecasting horizons longer than one, indicating its suitability for longer-term forecasts. This advantage arises from the inclusion of copulas, which effectively capture the dependence structure of residuals, thereby enhancing forecasting accuracy.

Some may observe that the hierarchical model without copulas performs worse than MinT(Shrink). This outcome stems from the selected top-down structure, as recommended in Hyndman et al. (2011). Nonetheless, this structure enables the model to capture the hierarchical nature of the data and integrate exogenous information related to house prices.

While the choice between the models depends on the research question and the nature of the data being analysed, the hierarchical model's numerical approach can offer a promising alternative to more complex modelling techniques in the field.

4.4 Comparing the hierarchical model and the FAVAR model

This subsection compares the hierarchical and FAVAR models, excluding residual analysis, in terms of goodness of fit, model complexity, and prediction accuracy. The FAVAR methodology is a potential factor model for addressing the limited information problem, as it integrates standard structural VAR analysis with dynamic factor analysis for large datasets (Bernanke et al., 2005). Dynamic factor models suggest that information from numerous observations can be summarised by a relatively small set of estimated indexes or latent factors.

Three subtypes of hierarchical models and the two FAVAR models are compared from the perspectives of model construction and interpretability, and the comparison is shown in Table 5. These three subtypes are simplified hierarchical, half hierarchical, and full hierarchical models, with the simplified hierarchical model already introduced in Subsection 4.3. The moving average terms are excluded to maintain consistency with the other two hierarchical models. The half-hierarchical model refers to a model whose risk factors are constructed at the SA4 level according to the special case 1 of the RP-PCA model as expressed in Equation (18).

The reason for comparing the FAVAR and the half/full hierarchical models is that the final prediction model for each area in the FAVAR and the hierarchical models contain similar components, consisting of 13 risk factors. In the first type of FAVAR, denoted as FAVAR1, and the hierarchical models, these risk factors can vary geographically at the same time point, whereas the risk factors in the second type of FAVAR model, denoted as FAVAR2, are the same.[14] The prediction accuracy of the training data represents the goodness of fit for different models. From Table 5, it is observed that an increase in model complexity leads to an increase in the degree of data interpretation. Comparing the half and full hierarchical models indicates that the use of the RP-PCA model significantly increases the model's interpretability.

[14]The introduction and constructions of FAVAR models are given in Appendix D

Table 5. Comparing FAVAR models and hierarchical models for goodness of fit.

Model	National Variables	SA4-Level Variables	National Factors (No.)	SA4-Level Factors (No.)	Accuracy ($\times 10^{-2}$)
FAVAR1	Yes	No	$(6+1)\times\|\Omega\|$	$(0+6)\ \times\|\Omega_j\|$	1.354
FAVAR2	Yes	No	$(6+7)\times\|\Omega\|$	$0\ \times\|\Omega_j\|$	1.868
Hierarchical(S)	No	No	$(0+1)\ \times\|\Omega\|$	$(0+1)\ \times\|\Omega_j\|$	1.903
Hierarchical(H)	Yes	Yes	$(6+1)\ \times\|\Omega\|$	$(5+1)\times\|\Omega_j\|$	1.626
Hierarchical(F)	Yes	Yes	$(6+1)\ \times\|\Omega\|$	$(5+1)\times\|\Omega_j\|$	1.421

Notes: The prediction accuracy is presented by the average training RMSE of different rolling windows with a length of 120. The table compares the performance of two FAVAR models, FAVAR1 and FAVAR2, along with three hierarchical models - simplified, half, and full. In the columns of national and SA4-Level risk factors, the expression $(a+b)\times c$ signifies the sum of a risk factors that are derived from external data sources, added to b risk factors that are derived from the house price data itself, and multiplied by the total number of areas in that particular hierarchy c.

Table 6 provides evidence that incorporating more external risk factors into the hierarchical model does not significantly compromise the accuracy of future forecasts. We notice that although the full hierarchical model does not perform the best in the short-term forecasting period, its average RMSE becomes more consistent across areas at the postcode level as the forecasting time increases. The FAVAR1 model, despite having a similar format and complexity as the full hierarchical model, performs the worst in long-term forecasting due to its high model complexity. It is also worth noting that the full hierarchical model outperforms the half-hierarchical model in shorter and longer forecasting periods. This suggests that introducing external data through RP-PCA can help reduce errors in long-term forecasting. In contrast, the FAVAR2 model used for comparison exhibits superior predictive performance. One possible reason is that the FAVAR2 model is predictively preferred, which leads to good forecasting performance but limited interpretability. This viewpoint is also supported by the results in Table 5.

Table 6. Comparative table of RMSE for one to twelve steps ahead predictions using factor models.

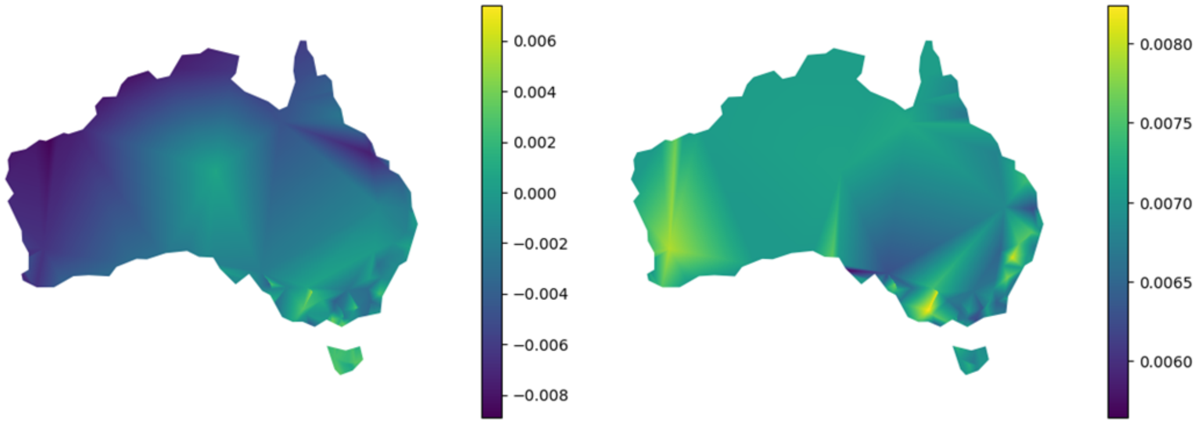
	$h=1$	1–3	1–6	1–9	1–12
Hierarchical(S) ($\times 10^{-2}$)	2.34	2.53	2.59	2.62	2.61
FAVAR1 (%)	-3.85	-0.67	1.74	3.45	4.60
FAVAR2 (%)	-4.49	-2.34	-0.93	-0.34	-0.55
Hierarchical(H)(%)	-3.72	-0.41	1.61	1.96	2.29
Hierarchical(F)(%)	-4.28	-0.94	0.32	1.79	1.86

Notes: The hierarchical(S) line shows the average RMSE ($\times 10^{-2}$) of the simplified hierarchical models based on AR models. A negative (positive) entry above this row shows the percentage decrease (increase) in the average RMSE of forecasts relative to the base forecasts.

Overall, this subsection compares two FAVAR models and three hierarchical models, and Table 6 supports adopting the RP-PCA model within the hierarchical framework, affirming its credibility in offering enhanced interpretability and decent forecasts for HPIs.

4.5 Analysis of risk factors

This subsection explores potential factors explaining the variation in historical house prices across different suburbs and highlights key findings from the model clarifying these factors.



(a) $\beta_i^P(0)$ **(b)** $\mathrm{C}_l(\Omega_{ji})$

Figure 7. Illustration of HPI growth rates as modelled by the hierarchical one-step-ahead forecasting model. Subplot 7(a) highlights the differences in fixed growth rates, indicated by the constant term $\beta_i^P(0)$ from Proposition 3.3, across various suburbs. Additionally, Subplot 7(b) demonstrates the influence of the national housing market, as evidenced by the coefficients of l_{t-1}, $\mathrm{C}_l(\Omega_{ji})$, in each suburb.

The final hierarchical one-step-ahead forecasting model is represented as a linear combination of a constant term, the current national-level risk factors, and the current SA4-level risk factors, including l_t and $l_{j,t}$. All risk factors are expressed as a linear combination of those from previous periods by substituting them with past factors using the joint dynamics in Proposition 3.3.

Figure 7(a) shows that house prices in different areas have varying fixed growth rates, represented by the final model's constant term $\beta_i^P(0)$ in Proposition 3.3. Suburbs closer to the southeast and big cities of Australia have higher fixed growth rates. In comparison, suburbs closer to the northwest and away from big cities have lower fixed growth rates. Furthermore, the national housing market significantly impacts certain areas, quantified by the coefficients of l_{t-1} in each suburb, denoted as $\mathrm{C}_l(\Omega_{ji})$. As shown in Figure 7(b), HPIs in metropolitan areas have experienced growth on average during these years, largely driven by the overall housing market in the previous term.[15]

At the SA4 level, five exogenous risk factors exist, and one l_j is the distance of fixed house price growth rates between the national and SA4 levels. Despite employing a consistent construction method, the coefficients and values associated with these risk factors vary. Consequently, a sole analysis of coefficients may not suffice in comprehensively evaluating the impact of these risk factors. One idea is that the sum of the partial effects of historical changes of risk factors at the SA4 level is considered the contribution of risk factors to HPI changes in each area.

To assess the impact of a specific factor f on house price changes, we calculate the ratio of its annual

[15]The coefficients of risk factors except for l_t at the national level, along with the corresponding geographical distribution, are presented in Appendix E. A more in-depth analysis of these risk factors will be provided in conjunction with the SA4-level risk factors in this subsection.

change effect to the impact of the annual change in the log of national HPI growth:[16]

$$\frac{\mathrm{C}_f\left(\Omega_{ji}\right) \times \sum_{t<T} f_t^P\left(\Omega_j\right)}{\mathrm{C}_l\left(\Omega_{ji}\right) \times \sum_{t<T} l_t}, \tag{23}$$

which is considered the effect of the factor on house price changes for time T. The coefficient $\mathrm{C}_f\left(\Omega_{ji}\right)$ still refers to the one obtained after expanding the final observation based on the joint dynamics concerning the specific risk factor f.[17] This constructed measure also distorts the average house price caused by a particular factor at the SA4 level in a suburb. As these risk factors (excluding l and l_j) are linear combinations of variables that have already been centralised, standardised, and made stationary, the partial effects of historical changes in risk factors at the SA4 level are expected to converge towards zero in the long run. The historical dataset supports this expectation, demonstrated in Figure 8, which displays the trend of average partial effects of the factors in the model over the years in all suburbs.

Figure 8 shows that the average partial effects of historical changes in l_j become gradually stable at a relatively constant level over time. In addition, although the partial effects of SA4-level risk factors for a one-period change may be significant, as time progresses, the average partial effects of historical changes in most areas rapidly converge towards zero and then slightly oscillate in a minimal range. Therefore, in long-term future forecasts, the cumulative effects of l_j terms will still be the most significant. At the same time, other risk factors at the SA4 level will only have a cumulative impact in the short term. This aligns with the intention to build a hierarchical model, which aims to provide reasonable explanations in one-step-ahead prediction without compromising long-term forecasting results.

4.6 Sensitivity analysis of the final model

In this subsection, we vary one single preprocessed variable from Table 2, with the remaining variables fixed. The change is assumed to increase one standard deviation from the initial value. Initial values for variables at the SA4 level are incremented by an amount equal to their corresponding standard deviations. We observe how this one-unit change will influence the predicted HPI (average) in 5 years across different areas on the postcode level. Table 7 illustrates the average and standard deviation of predicted house price changes across different areas. The impact of changes is notably more significant for responses with a higher mean value. In contrast, the effects of the changes are more volatile across different suburbs for responses with a larger standard deviation.

[16]The reason for not using the annual changes in the local housing market change in each suburb as the denominator is that sometimes the absolute values of changes in the overall market can be very small in each suburb, leading to an overestimation of the effect. However, the sum of annual l tends to remain at a relatively stable level, with the coefficient of l all positive as illustrated in Table 13, Appendix E, making it a more suitable denominator for comparison.

[17]As shown in Figure 12, Appendix E, one risk factor's contributions vary significantly across different areas. However, further explanations cannot be provided due to the lack of clear economic interpretations of these risk factors. Additionally, attempting to analyse the original variables of these risk factors is not preferred, considering that the principal component of the analysis is to eliminate the possible correlations between different variables and reduce dimensionality.

Table 7. Changes in the average and standard deviation of predicted house prices across areas.

Description	Mean	Standard deviation
Interest rate	-2.23%	3.76%
Exchange rate	-1.74%	6.16%
Consumer price index	-2.01%	2.62%
Gross domestic product	5.93%	3.56%
Retail sales	2.88%	4.29%
Private dwelling approvals	3.25%	4.98%
Australian securities exchange	3.10%	2.72%
Total home debt	0.32%	4.72%
Current weekly gross wages & salary	3.96%	8.23%
Home repairs/renovations/maintenance	3.32%	1.65%
Fees paid to health practitioners	4.84%	7.16%
Weekly time on paid employment	3.12%	6.86%
Went without meals	-2.26%	7.19%
Could not pay the mortgage or rent on time	-5.41%	2.13%
Employment status	3.12%	1.45%
Median of age	-0.77%	1.02%
The standard deviation of age	2.32%	4.32%
The third quantile of age	-0.18%	2.01%
Total number of people	2.62%	1.90%
Country of birth: Australia	0.08%	2.18%
Country of birth: New Zealand	0.35%	1.11%
Country of birth: India	-0.23%	3.86%
Country of birth: China (excludes SARs)	0.52%	2.92%
Country of birth: Vietnam	0.36%	3.41%
Aboriginal	0.15%	4.21%
Torres Strait Islander	-0.24%	1.99%
Highest education level: Year 12 or equivalent	0.41%	4.42%
Highest education level: Certificate III/IV	1.62%	4.10%
Highest education level: Bachelor degree	2.12%	3.29%
Highest education level: Postgraduate degree	2.36%	3.98%

Notes: Table 7 presents the impact of a one standard deviation increase in a selected preprocessed variable from Table 2, holding all other variables constant. This analysis aims to understand the influence of one single variable adjustment on the predicted average HPI over five years at the postcode level. The table summarises the average and standard deviation of these predicted changes in house prices across various areas, providing insights into the sensitivity of HPIs to specific variable alterations in different locations. The dashed lines separate the table into three parts: national-level variables, economic variables at the SA4 level, and demographic variables at the SA4 level, differing from the separation criteria used in Table 2.

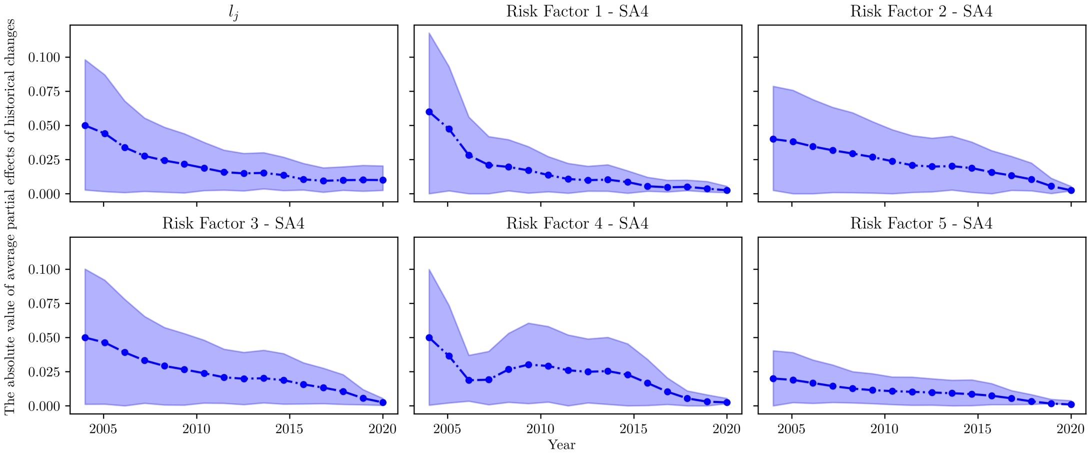


Figure 8. Partial effects of historical changes in SA4-level risk factors. This figure shows the absolute values of the average partial effects of historical changes over time in all suburbs. It illustrates the decreasing trend of absolute values in the average historical changes of risk factors over time. The line represents the mean value across all suburbs, while the interval encompasses 99% of the suburbs.

We start with a discussion of the national-level variables. An increase in interest rates elevates borrowing costs and decreases house demand. This effect will be more pronounced in suburbs located in metropolitan areas such as Sydney and Melbourne.[18] Exchange rate movements also play a significant role. A higher appreciation of the local currency relative to foreign currencies results in a decline in foreign investment in the housing sector. Suburbs with considerable foreign investors, such as Sydney CBD, Pyrmont, Chatswood in New South Wales, Melbourne CBD and Box Hill in Victoria, and Broadbeach in Queensland are anticipated to be particularly sensitive to exchange rates. When not complemented by corresponding wage hikes, a surge in the CPI will generally suppress house demand across all suburbs. GDP growth will uplift house prices, reflecting the nation's economic robustness. This effect is directly observable in GDP-centric areas in New South Wales, such as North Sydney and Parramatta, and those in Victoria, such as Docklands and Clayton. Retail sales growth signals a potential increase in housing market activity. Commercial hubs such as CBD areas in cities will witness a corresponding surge in house prices. Lastly, the performance of the Australian Securities Exchange plays its part. A bullish swing in the ASX will spur the real estate market. Wealthier suburbs, including Point Piper and Bellevue Hill in New South Wales, Toorak, South Yarra, East Melbourne in Victoria, and Ascot and Hamilton in Queensland, will likely exhibit this trend more prominently.

Next, we analyse the economic variables at the SA4 level. A rise in home debt signifies two concurrent phenomena: increased buying activity, which pushes house prices up (among suburbs in Sydney and Melbourne), or burgeoning financial stress, which will later depress house prices (natural towns or natural landscapes) (Mian and Sufi, 2011). Wage growth typically bolsters purchasing power, potentially increasing house prices. Suburbs with high employment rates in major cities will be especially responsive. However, suburbs with low employment rates may not experience an appreciation in house prices. These areas may not experience a significant rise in purchasing power among residents. In some cases, if the broader mar-

[18]The maps for metropolitan areas in Sydney, Melbourne, Brisbane, and Perth are provided in Appendix F.

ket is experiencing overall growth in purchasing power but these areas are not, it might lead to a relative depreciation in house prices. A surge in home improvement activities generally leads to housing value appreciation with minor geographical variations. Fees paid to health practitioners serve as an indicator for health-conscious locations such as Bellevue Hill and Vaucluse in New South Wales, Toorak in Victoria, and Ascot in Queensland, where a greater increase is associated with more expensive real estate. If employment hours increase, it will reflect economic vitality, correspondingly boosting housing demand and prices, particularly in commercial hubs. A spike in financial distress indicators ("Went without meals" or "Could not pay the mortgage or rent on time") foreshadows future price corrections, with the most vulnerable areas, such as parts of Western Sydney and outer suburbs of Melbourne, potentially showing sharper downturns.[19] A shift in employment implies higher disposable income levels and borrowing capacities, generally pushing house prices. In the Australian context, the influx of immigrants further contributes to these trends by increasing urban populations, thereby intensifying competition for housing and sustaining upward price pressure.

We also examine the demographic-related variables at the SA4 levels. Population growth usually drives an increase in house prices, and areas with high population density, such as Surry Hills and Chippendale in New South Wales, Southbank and Fitzroy in Victoria, Kangaroo Point in Queensland, East Perth in Western Australia, and Bowden in South Australia, often encounter more pronounced housing supply limitations, exacerbating the rise in house prices. Interestingly, greater variation in age distribution tends to elevate house prices, particularly in major cities. This suggests that a balanced population structure would contribute to increased property values. With a surge in higher education attainments (Year 12, Certificate III/IV[20], Bachelor degree, and Postgraduate degree), a burgeoning house market will be observed due to upward shifts in income brackets and housing preferences. Other variables, such as the country of birth and indigenous status, may suggest housing needs and financial capabilities. However, they have smaller influences than other variables, without obvious trends observed about the variation across different areas.

This sensitivity analysis explores the intricate relationships between variables at different geographical levels and house prices across different areas. Notably, the results underscore the imperative of examining risk factors at more granular levels. Ignoring the granularity can obscure significant nuances and lead to generalised decisions that might not apply uniformly across all areas. Hence, to ensure a comprehensive understanding of house prices and to devise effective strategies related to home equity, it is necessary to incorporate and analyse risk factors at these detailed levels.

[19]These suburbs often exhibit lower household incomes and reduced access to high-paying employment opportunities, contributing to their financial fragility.

[20]Certificates III-IV provide training in more advanced skills and knowledge. Universities generally accept a Certificate IV equivalent to six to twelve months of a bachelor degree.

5 Conclusion

In the existing literature describing and forecasting local house prices using the VAR model and its extensions, the vector autoregressive model with exogenous variables (VARX) model introduced in Shao et al. (2015) remains the sole approach employing granular-level information dynamically to explain the geographical variation of house price growth rates. However, only house prices are featured as risk factors in their model, which seems reasonable, given that other cumulative partial effects of risk factors at granular levels have been shown to converge to zero in the long run. Nevertheless, this paper demonstrates that, in the short term, other risk factors at granular levels substantially affect forecasts. To incorporate these influences, this paper, drawing on the conditional CAPM proposed in Jagannathan and Wang (1996), constructs a hierarchical structure that provides a robust basis for incorporating external variables from various areas. To keep the model manageable, we adopt the spatial application of the RP-PCA method from Lettau and Pelger (2020), consolidating external variables into newly defined lower-level risk factors for integration into the model. Using information from a model's residual covariance structure has enhanced prediction accuracy, as evidenced in studies such as Wickramasuriya et al. (2018). Nevertheless, the residuals derived from the hierarchical model do not follow a normal distribution, rendering theoretical estimates of the adjusted model challenging to obtain. We propose a simulation-based approach incorporating copulas of these residuals into the hierarchical model.

The contribution of this paper lies in providing a method to incorporate lower-level risk factors without significantly impairing the long-term predictability of the model. The conditional CAPM, RP-PCA, and residual analysis jointly contribute to this preservation of predictability. Introducing these risk factors has enhanced the model's short-term prediction and interpretability. At the same time, over a longer period, the cumulative effects of these introduced factors gradually tend towards zero, ensuring that the model's long-term predictive performance is not significantly compromised. As a result, this model can be used as a reference for further analysis of reverse mortgages based on house prices. This helps to more comprehensively assess the risks and returns of reverse mortgage products, providing decision-makers with additional information on house price risks.

Future research could explore variant coefficients in the suburbs to capture evolving house price risk more accurately, thereby improving the model's predictive performance. Moreover, the current residual analysis only considers one-step-ahead residuals representing both temporal and spatial dimensions, as some theoretical derivations in this paper shift conclusions from time to space. Consequently, further investigation is needed to determine whether the final residuals are related to the residuals produced between different levels in the hierarchical model, given their potential influence on the model's stability and predictability. Examining the influence of multi-step residuals may also help elucidate and address the co-dependence structure of house price changes across areas.

Furthermore, this paper notes that HPI data typically cluster near coastal areas and are strongly influenced by metropolitan areas situated far apart in the Australian context. However, extending this framework to other countries requires accommodating diverse geographical, demographic, and socio-economic features. Consequently, future studies should incorporate aspects such as population density, connectivity between areas, and geographical characteristics to identify suitable area classifications.

Acknowledgements

The authors acknowledge financial support from the UNSW Business School, the Australian Research Council Centre of Excellence in Population Ageing Research, and the Australian Research Council Discovery Project: DP210101195.

This paper uses unit record data from Household, Income and Labour Dynamics in Australia Survey conducted by the Australian Government Department of Social Services (DSS). However, the findings and views reported in this paper are those of the author(s) and should not be attributed to the Australian Government, DSS, or any of DSS' contractors or partners. DOI: 10.26193/PI5LP.

The authors would like to acknowledge and express gratitude to CoreLogic for providing house price index data.[21] Any interpretations, conclusions, or potential errors within this paper are solely those of the authors and are not representative of CoreLogic's views or positions.

[21]For further details on the dataset, please refer to the CoreLogic website at https://www.corelogic.com.au/.

References

Actuaries Institute (2024). More than just a roof: Changing the narrative on the role of the home. Dialogue paper. Available on: https://www.actuaries.asn.au/docs/thought-leadership-reports/more-than-just-a-roof.pdf?Status=Temp&sfvrsn=f370bef4_6.

Adams, Z. and Füss, R. (2010). Macroeconomic determinants of international housing markets. *Journal of Housing Economics*, 19(1):38–50.

Aquaro, M., Bailey, N., and Pesaran, M. H. (2021). Estimation and inference for spatial models with heterogeneous coefficients: An application to US house prices. *Journal of Applied Econometrics*, 36(1):18–44.

Atalay, K. and Edwards, R. (2022). House prices, housing wealth and financial well-being. *Journal of Urban Economics*, 129:103438.

Australian Bureau of Statistics (2023). Total value of dwellings. https://www.abs.gov.au/statistics/economy/price-indexes-and-inflation/total-value-dwellings/latest-release.

Baltagi, B. H. and Li, J. (2014). Further evidence on the spatio-temporal model of house prices in the United States. *Journal of Applied Econometrics*, 29(3):515–522.

Banasiak, L. (2016). Yahoo-finance. https://github.com/yahoo-finance/yahoo-finance.

Basile, R., Durbán, M., Mínguez, R., Montero, J. M., and Mur, J. (2014). Modeling regional economic dynamics: Spatial dependence, spatial heterogeneity and nonlinearities. *Journal of Economic Dynamics and Control*, 48:229–245.

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly Journal of Economics*, 120(1):387–422.

Bhattacharjee, A., Castro, E., Maiti, T., and Marques, J. (2016). Endogenous spatial regression and delineation of submarkets: A new framework with application to housing markets. *Journal of Applied Econometrics*, 31(1):32–57.

Bin, O. (2004). A prediction comparison of housing sales prices by parametric versus semi-parametric regressions. *Journal of Housing Economics*, 13(1):68–84.

Bourassa, S. C., Hoesli, M., and Peng, V. S. (2003). Do housing submarkets really matter? *Journal of Housing Economics*, 12(1):12–28.

Brady, R. R. (2011). Measuring the diffusion of housing prices across space and over time. *Journal of Applied Econometrics*, 26(2):213–231.

Caldera, A. and Johansson, Å. (2013). The price responsiveness of housing supply in OECD countries. *Journal of Housing Economics*, 22(3):231–249.

Chomik, R. and Yan, S. (2019). Housing in an ageing Australia: Nest and nest egg? Research Brief, Centre of Excellence in Population Ageing Research (CEPAR). Available on: https://cepar.edu.au/resources-videos/research-briefs/housing-ageing-australia-nest-and-nest-egg.

CoreLogic (2024). CoreLogic indices. https://www.corelogic.com.au/our-data/corelogic-indices#daily-indices.

Dunn, D. M., Williams, W. H., and DeChaine, T. (1976). Aggregate versus subaggregate models in local area forecasting. *Journal of the American Statistical Association*, 71(353):68–71.

Fama, E. F. and MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81(3):607–636.

Fève, P. and Moura, A. (2024). Frictionless house-price momentum. *Journal of Economic Dynamics and Control*, 168:105000.

Ganduri, R. and Maturana, G. (2024). Do property rehabs affect neighboring property prices? *Journal of Urban Economics*, 143:103694.

Gevorgyan, K. (2019). Do demographic changes affect house prices? *Journal of Demographic Economics*, 85(4):305–320.

Ghysels, E. (2016). Macroeconomics and the reality of mixed frequency data. *Journal of Econometrics*, 193(2):294–314.

Glaeser, E. L. and Gyourko, J. (2018). The economic implications of housing supply. *Journal of Economic Perspectives*, 32(1):3–30.

Goodhart, C. and Hofmann, B. (2008). House prices, money, credit, and the macroeconomy. *Oxford Review of Economic Policy*, 24(1):180–205.

Goodman, A. C. and Thibodeau, T. G. (2003). Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics*, 12(3):181–201.

Goodman, A. C. and Thibodeau, T. G. (2008). Where are the speculative bubbles in US housing markets? *Journal of Housing Economics*, 17(2):117–137.

Google (2024). Google Earth. https://www.google.com/earth/about/.

Granziera, E. and Kozicki, S. (2015). House price dynamics: Fundamentals and expectations. *Journal of Economic Dynamics and Control*, 60:152–165.

Green, R. K. and Malpezzi, S. (2001). *A Primer on US Housing Markets and Housing Policy*. Urban Institute Press.

Gross, C. W. and Sohl, J. E. (1990). Disaggregation methods to expedite product line forecasting. *Journal of Forecasting*, 9(3):233–254.

Guerrieri, V., Hartley, D., and Hurst, E. (2013). Endogenous gentrification and housing price dynamics. *Journal of Public Economics*, 100:45–60.

Hanewald, K. and Sherris, M. (2013). Postcode-level house price models for banking and insurance applications. *Economic Record*, 89(286):411–425.

Hilber, C. A. and Vermeulen, W. (2016). The impact of supply constraints on house prices in England. *The Economic Journal*, 126(591):358–405.

Ho, A. T., Huynh, K. P., and Jacho-Chávez, D. T. (2016). Flexible estimation of copulas: An application to the US housing crisis. *Journal of Applied Econometrics*, 31(3):603–610.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., and Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, 55(9):2579–2589.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(3):1–22.

Hyndman, R. J., Lee, A. J., and Wang, E. (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics & Data Analysis*, 97:16–32.

Jagannathan, R. and Wang, Z. (1996). The conditional CAPM and the cross-section of expected returns. *The Journal of Finance*, 51(1):3–53.

Kuttner, K. and Shim, I. (2012). Taming the real estate beast: The effects of monetary and macroprudential policies on housing prices and credit. RBA Annual Conference, Reserve Bank of Australia. Available on: https://www.rba.gov.au/publications/confs/2012/kuttner-shim.html.

Lee, Y., Kemp, P. A., and Reina, V. J. (2022). Drivers of housing (un)affordability in the advanced economies: A review and new evidence. *Housing Studies*, 37(10):1739–1752.

Lettau, M. and Pelger, M. (2020). Estimating latent asset-pricing factors. *Journal of Econometrics*, 218(1):1–31.

Liu, B., Mavrin, B., Niu, D., and Kong, L. (2016). House price modeling over heterogeneous regions with hierarchical spatial functional analysis. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 1047–1052. IEEE.

Lutkepohl, H. and Kratzig, M. (2004). *Time Series Econometrics.* Springer.

Mankiw, N. G. and Weil, D. N. (1989). The baby boom, the baby bust, and the housing market. *Regional Science and Urban Economics*, 19(2):235–258.

Melbourne Institute (2019). HILDA statistical report. Report. Available on: https://melbourneinstitute.unimelb.edu.au/__data/assets/pdf_file/0010/3398464/HILDA-Statistical-Report2019.pdf.

Mian, A. and Sufi, A. (2011). House prices, home equity–based borrowing, and the US household leverage crisis. *American Economic Review*, 101(5):2132–2156.

Minetti, R., Peng, T., and Jiang, T. (2019). Keeping up with the Zhangs and house price dynamics in China. *Journal of Economic Dynamics and Control*, 109:103778.

My Aged Care (2024a). Aged care home costs and fees. https://www.myagedcare.gov.au/aged-care-home-costs-and-fees.

My Aged Care (2024b). Home Care Package costs and fees. https://www.myagedcare.gov.au/help-at-home/commonwealth-home-support-programme.

Nystrup, P., Lindström, E., Pinson, P., and Madsen, H. (2020). Temporal hierarchies with autocorrelation for load forecasting. *European Journal of Operational Research*, 280(3):876–888.

Orcutt, G. H., Watts, H. W., and Edwards, J. B. (1968). Data aggregation and information loss. *American Economic Review*, 58(4):773–787.

Pace, R. K., Barry, R., Gilley, O. W., and Sirmans, C. (2000). A method for spatial–temporal forecasting with an application to real estate prices. *International Journal of Forecasting*, 16(2):229–246.

Qian, W. (2023). House price expectations and household consumption. *Journal of Economic Dynamics and Control*, 151:104652.

Reserve Bank of Australia (2023). Statistical tables. https://www.rba.gov.au/statistics/tables/.

Saiz, A. (2010). The geographic determinants of housing supply. *The Quarterly Journal of Economics*, 125(3):1253–1296.

Service Australia (2024). Who can get it. https://www.servicesaustralia.gov.au/who-can-get-age-pension?context=22526.

Shao, A. W., Hanewald, K., and Sherris, M. (2015). Reverse mortgage pricing and risk analysis allowing for idiosyncratic house price risk and longevity risk. *Insurance: Mathematics and Economics*, 63:76–90.

Shen, S. and Pang, J. (2018). Measuring the diffusion of housing prices across space and over time: Replication and further evidence. *Journal of Applied Econometrics*, 33(3):479–484.

Shi, S., Jou, J.-B., and Tripe, D. (2014). Can interest rates really control house prices? Effectiveness and implications for macroprudential policy. *Journal of Banking & Finance*, 47:15–28.

Takáts, E. (2012). Aging and house prices. *Journal of Housing Economics*, 21(2):131–141.

Titman, S. and Zhu, G. (2024). City characteristics, land prices and volatility. *Journal of Urban Economics*, 140:103645.

Wang, E. (2021). Hierarchical and grouped time series. https://github.com/earowang/hts.

Wetzstein, S. (2017). The global urban housing affordability crisis. *Urban Studies*, 54(14):3159–3177.

Wickramasuriya, S. L., Athanasopoulos, G., and Hyndman, R. J. (2018). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, 113(522):394–404.

Wrenn, D. H., Klaiber, H. A., and Newburn, D. A. (2017). Confronting price endogeneity in a duration model of residential subdivision development. *Journal of Applied Econometrics*, 32(3):661–682.

Zhu, B., Betzinger, M., and Sebastian, S. (2017). Housing market stability, mortgage market structure, and monetary policy: Evidence from the euro area. *Journal of Housing Economics*, 37:1–21.

Appendix A Proofs

Proof of Proposition 3.1:

Proof. If the information set $I_{i,t-1}(\Omega)$ is known, the first $\varkappa$ partial risk factors and premia are estimable. In this case,

$$\begin{aligned}
&\mathbb{E}\left[h_{i,t}^P \mid I_{i,t-1}\left(\Omega_j, \Omega_{ji}\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[h_{i,t}^P \mid I_{i,t-1}\left(\Omega, \Omega_j, \Omega_{ji}\right)\right] \mid I_{i,t-1}\left(\Omega_j, \Omega_{ji}\right)\right] \\
&= \mathbb{E}\left[\mathbf{w}_i^P \mid I_{i,t-1}\left(\Omega_j, \Omega_{ji}\right)\right]^\top \mathbb{E}\left[\boldsymbol{f}_{i,t}^P \mid I_{i,t-1}\left(\Omega_j, \Omega_{ji}\right)\right] \\
&\quad + \mathbb{C}\mathrm{ov}_\omega\left(\mathbb{E}\left[\mathbf{w}_i^P \mid I_{i,t-1}\left(\Omega, \Omega_j, \Omega_{ji}\right)\right], \mathbb{E}\left[\boldsymbol{f}_{i,t}^P \mid I_{i,t-1}\left(\Omega, \Omega_j, \Omega_{ji}\right)\right] \mid I_{i,t-1}\left(\Omega_j, \Omega_{ji}\right)\right),
\end{aligned} \tag{24}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{w}_i^P \mid I_{i,t-1}\left(\Omega_j, \Omega_{ji}\right)\right]^\top \mathbb{E}\left[\boldsymbol{f}_{i,t}^P \mid I_{i,t-1}\left(\Omega_j, \Omega_{ji}\right)\right] \\
&= \mathbf{w}_i^P([1:\varkappa])^\top \boldsymbol{f}_{i,t}^P([1:\varkappa]) \\
&\quad + \mathbb{E}\left[\mathbf{w}_i^P(-[1:\varkappa]) \mid I_{i,t-1}\left(\Omega_j, \Omega_{ji}\right)\right]^\top \mathbb{E}\left[\boldsymbol{f}_{i,t}^P(-[1:\varkappa]) \mid I_{i,t-1}\left(\Omega_j, \Omega_{ji}\right)\right].
\end{aligned} \tag{25}$$

□

Proof of Lemma 3.4 and Remark 3.2:

Proof. According to Proposition 3.1, if previous risk factors are considered to be influential on house prices, the basic HPI model at the national level is

$$\text{HPI of suburb } i: h_{i,t}^P(\Omega) = \mathbf{w}_i^P(\Omega)^\top \boldsymbol{f}_{i,t}^P(\Omega) \sum_{\kappa>0} L^\kappa + \eta_t, \tag{26}$$

where $\mathbf{w}_i^P(\Omega)$ is a vector of regression coefficients illustrating partial risk premia about national-level risk factors $\boldsymbol{f}_{i,t}^P(\Omega)$ in the previous periods, L is the lag operator, and η_t represents a random residual with finite mean and variance, which can be expressed as a combination of current and past random disturbance terms. The specific form of η_t will be determined in the model selection process. It is worth noting that the national-level risk factors $\boldsymbol{f}_{i,t}^P(\Omega)$ can be written as $\boldsymbol{f}_t^P(\Omega)$ because these risk factors are assumed to be the same across different areas at a certain time point. Therefore, the relationship between the real house price growth index and the predicted HPI at the national level is

$$\begin{aligned}
h_{i,t} &= h_{i,t}^P(\Omega) + e_{i,t}(\Omega) + \psi_t \\
&= \bar{\mathbf{w}}_i^P(\Omega)^\top \boldsymbol{f}_t^P(\Omega) \sum_{\kappa>0} L^\kappa + \left(\mathbf{w}_i^P(\Omega)^\top - \bar{\mathbf{w}}_i^P(\Omega)^\top\right) \boldsymbol{f}_t^P(\Omega) \sum_{\kappa>0} L^\kappa + \eta_t + e_{i,t}(\Omega) + \psi_t \\
&= \bar{\mathbf{w}}_i^P(\Omega)^\top \boldsymbol{f}_t^P(\Omega) \sum_{\kappa>0} L^\kappa + \boldsymbol{\beta}_i^P(\Omega)^\top \boldsymbol{f}_t^P(\Omega) \sum_{\kappa>0} L^\kappa + \eta_t + e_{i,t}(\Omega) + \psi_t,
\end{aligned} \tag{27}$$

where $\bar{\mathbf{w}}_i^P(\Omega)$ is the average of $\mathbf{w}_i^P(\Omega)$ and $\boldsymbol{\beta}_i^P(\Omega)$ is the deviation of $\mathbf{w}_i^P(\Omega)$ from the average, ψ_t is the error arising from the distribution-related assumption of η_t, and $e_{i,t}$ is the rest. Therefore, $e_{i,t}(\Omega) + \psi_t$ is the total residual of the one-level house price model for the i^{th} area. The expectation of η_t and $e_{i,t}$ are zero, while the specific distribution of them are unknown, which requires further analysis.

The residuals from the basic model are considered as the impact of unmeasurable risk factors and premia according to Proposition 3.1. Therefore, these residuals are assumed to be new excess returns, which will be analysed by models at smaller scales. Based on the basic model, more information is collected at a smaller scale; thus, more factors are introduced to the second level of the model:

$$\begin{aligned}
&\text{HPI of } i \in \Omega_j : h_{i,t}^P(\Omega_j) = l_t + l_{j,t} + \boldsymbol{\beta}_i^P(\Omega)^\top \boldsymbol{f}_t^P(\Omega) \sum_{\kappa>0} L^\kappa + \boldsymbol{\beta}_i^P(\Omega_j)^\top \boldsymbol{f}_t^P(\Omega_j) \sum_{\kappa>0} L^\kappa, \\
&\text{Average HPI: } l_t = \bar{\mathbf{w}}_i^P(\Omega)^\top \boldsymbol{f}_t^P(\Omega) \sum_{\kappa>0} L^\kappa + \eta_t, \\
&D(\Omega, \Omega_j) : l_{j,t} = \bar{\mathbf{w}}_i^P(\Omega_j)^\top \boldsymbol{f}_t^P(\Omega_j) \sum_{\kappa>0} L^\kappa + \eta_{j,t}.
\end{aligned} \tag{28}$$

Here all parameters and the scalar $\eta_{j,t}$ have similar definition as in the basic model, $D(\cdot,\cdot)$ measures the distance between HPIs at different levels, and $\boldsymbol{f}_{i,t}^P(\Omega_j)$ denotes SA4-level risk factors, which can also be written as $\boldsymbol{f}_t^P(\Omega_j)$ because these risk factors are assumed to be the same in the same area at the SA4 level. □

Proof of Proposition 3.2:

Proof. The lack of information can be interpreted as the lack of available data at each scale. According to Lemmas 3.1-3.3, and the introduction of lag terms, the difference between $\mathbb{E}[h_{i,t}]$ and $\mathbb{E}_\omega\left[h_{i,t}^P\right]$ is the error resulting from the lack of information. The final error of the hierarchical model is defined as

$$e_{i,t}(\Omega_j) = h_{i,t} - h_{i,t}^P - \psi_t - \psi_{j,t} = \mathbb{E}[h_{i,t}] + \tilde{\epsilon}_{i,t} - \left(\mathbb{E}_\omega[h_{i,t}^P] + \eta_t + \eta_{j,t}\right) - \psi_t - \psi_{j,t}. \tag{29}$$

Because VAR models are selected both at the national and SA4 levels, the residuals η_t (resp. $\eta_{j,t}$) are equivalent to the stochastic disturbances, which are noises with mean 0 and finite variance. □

Proof of Proposition 3.3

Proof. The first part of residuals represents time-variant means, including level/step shifts, seasonal pulses, and local time trends. The second part is regarded as a random term with time-varying variance. If $\mathbb{E}[h_{i,t}] - \mathbb{E}_\omega[h_{i,t}^P]$ is not time-variant in this model, it can be interpreted by a constant term $\boldsymbol{\beta}_i^P(0)$ in the final model.[22] This term only changes across different areas, illustrated as Equation (30):

$$\begin{aligned}&\text{HPI of suburb } i \in \Omega_j : h_{i,t}^P(\Omega_j) = l_t + l_{j,t} + l_{ji,t},\\ &D(\Omega_j, \Omega_{ji}) : l_{ji,t} = \boldsymbol{\beta}_i^P(0) + \boldsymbol{\beta}_i^P(\Omega)^\top \boldsymbol{f}_t^P(\Omega) \sum_{\kappa>0} L^\kappa + \boldsymbol{\beta}_i^P(\Omega_j)^\top \boldsymbol{f}_t^P(\Omega_j) \sum_{\kappa>0} L^\kappa.\end{aligned} \tag{30}$$

The only change from the previous model is adding the constant term, which differs in each suburb. Therefore, the final observation equation of the hierarchical model is

$$h_{i,t}^P(\Omega_j) = l_t + l_{j,t} + \boldsymbol{\beta}_i^P(0) + \boldsymbol{\beta}_i^P(\Omega)^\top \boldsymbol{f}_t^P(\Omega) + \boldsymbol{\beta}_i^P(\Omega_j)^\top \boldsymbol{f}_t^P(\Omega_j). \tag{31}$$

□

Appendix B Risk factors and joint dynamics

Table 8. Loadings of risk factors at the national level.

	PC1	PC2	PC3	PC4	PC5	PC6
ir	0.037548	-0.289941	0.380040	-0.747763	-0.448723	0.050139
exr	0.132970	-0.636969	0.065037	0.259925	0.126026	0.699006
cpi	0.564941	-0.037767	-0.097008	-0.292448	0.556420	-0.137104
gdp	0.683789	0.115184	-0.052727	-0.068571	-0.076643	0.036898
rs	0.421768	0.181933	0.398480	0.482083	-0.477737	-0.054874
pda	0.123707	-0.142536	-0.823083	-0.037965	-0.488915	0.020055
asx	0.031496	-0.664903	0.053827	0.221848	0.004148	-0.696638

Table 9. Information criteria for VARMA models with different lags at the national level.

	AIC	BIC	HQIC
(0,1)	3987.925	4276.723	4104.420
(0,2)	3978.774	4436.038	4163.225
(1,0)	3774.629	<u>4063.428</u>	<u>3891.125</u>
(1,1)	3741.851	4199.115	3926.302

	AIC	BIC	HQIC
(1,2)	3748.643	4374.373	4001.050
(2,0)	3751.895	4209.160	3936.346
(2,1)	<u>3720.842</u>	4346.573	3973.249
(2,2)	3752.779	4546.976	4073.142

The validation method tunes the penalty term, denoted by γ, over the range $[0, +\infty]$. For each value of γ, the number of risk factors $d_{90\%}$ can be identified, along with a consistent method for constructing risk factors across all SA4 structures. The number of risk factors $d_{90\%}$ is defined as the smallest number of risk factors explaining at least 90% of the variance. Figure 9 demonstrates that, as γ increases, $d_{90\%}$ gradually decreases.

[22] According to the Augmented Dickey-Fuller (ADF) test, which tests whether univariate time series data is stationary or not in the simulation, the proportion of series rejecting the hypothesis of being nonstationary is 0.9869 (454 out of 460) at the 1% significance level. Therefore, $\mathbb{E}[h_{i,t}] - \mathbb{E}_\omega[h_{i,t}^P]$ is not time-variant.

Table 10. Parameter estimates and covariance matrix in the VAR(1) model at national level.

Variable	PC1	PC2	PC3	PC4	PC5	PC6	l
				Parameter Estimates			
L1.PC1	0.025	-0.043	0.100	-0.112	0.077	-0.097	-0.234
L1.PC2	-0.088	0.020	0.077	0.079	0.080	0.018	0.013
L1.PC3	-0.112	-0.256	-0.337	0.058	-0.201	0.057	-0.060
L1.PC4	-0.009	0.031	-0.105	-0.350	-0.004	-0.086	-0.086
L1.PC5	0.084	-0.020	-0.188	0.070	-0.158	0.109	0.053
L1.PC6	-0.005	-0.018	0.019	-0.058	0.097	-0.020	0.017
L1.l	-0.072	-0.010	-0.026	0.033	-0.043	-0.049	0.924
			Cholesky Decomposition of Error Covariance				
PC1	1.155	-0.190	0.786	0.003	-0.244	0.899	0.037
PC2		0.214	0.074	0.863	-0.002	0.143	0.011
PC3			0.015	0.765	-0.078	-0.090	0.018
PC4				0.100	-0.026	0.709	-0.013
PC5					-0.027	-0.028	-0.043
PC6						0.003	-0.008
l							0.321

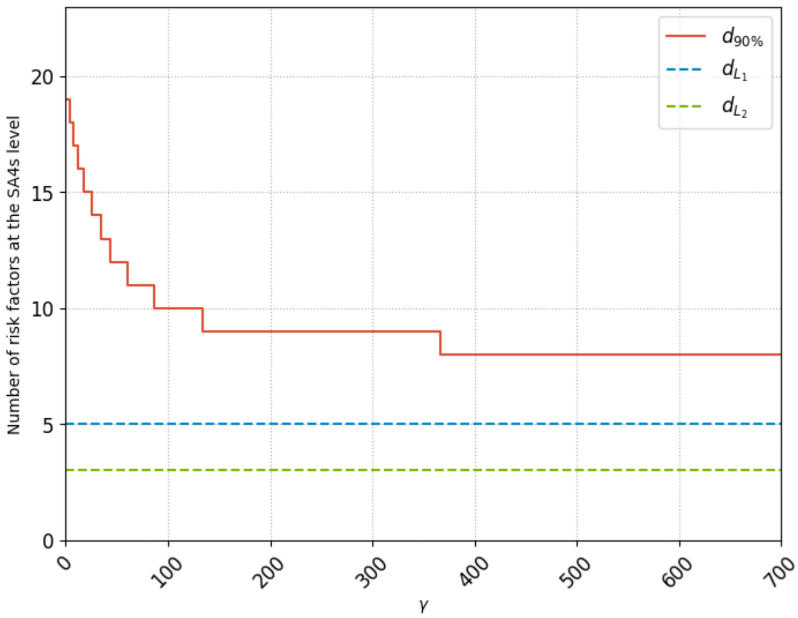


Figure 9. The relationship between the number of risk factors and penalty term γ in (21). The optimal selection of risk factors for each γ should consider the causal effect of risk factors on HPIs, measured by the RFE method, and the constraints represented by d_1 and d_2 dependent on the order of VAR models.

Upon exceeding a certain threshold, the number of principal components stabilises at 8, which corresponds to the maximal number of components obtained from principal component analysis of the cross-area average of variables according to Equation (18). However, incorporating all principal components as risk factors is not always the optimal approach, as it can increase the complexity of the model. Therefore, risk factors are ranked for each γ based on recursive feature elimination (RFE), a backward selection of the predictors. The risk factor that gets a higher average vote rate among all suburbs will rank higher.

The AIC of each VAR model based on the SA4-level risk factors is used to identify the optimal number of principal components. As the number of risk factors increases, the ability of the VAR model to explain the variation in multiple areas decreases due to the model's limitations, specifically, the inability to perform the Cholesky decomposition when the number of risk factors becomes too large. This characteristic narrows the search range for the optimal risk factors. Figure 9 illustrates the upper bounds of the number of risk factors using VAR models with lags 1 and 2, denoted as d_{L_1} and d_{L_2}. The increase in the lag parameter for VAR models at the SA4 level results in a downward shift of the upper bound of the maximal number of risk factors. This shows that even if PCA is used to reduce the dimensionality of the independent variables, it still cannot meet the requirement for successful use of the VAR model for each area at the SA4 level.

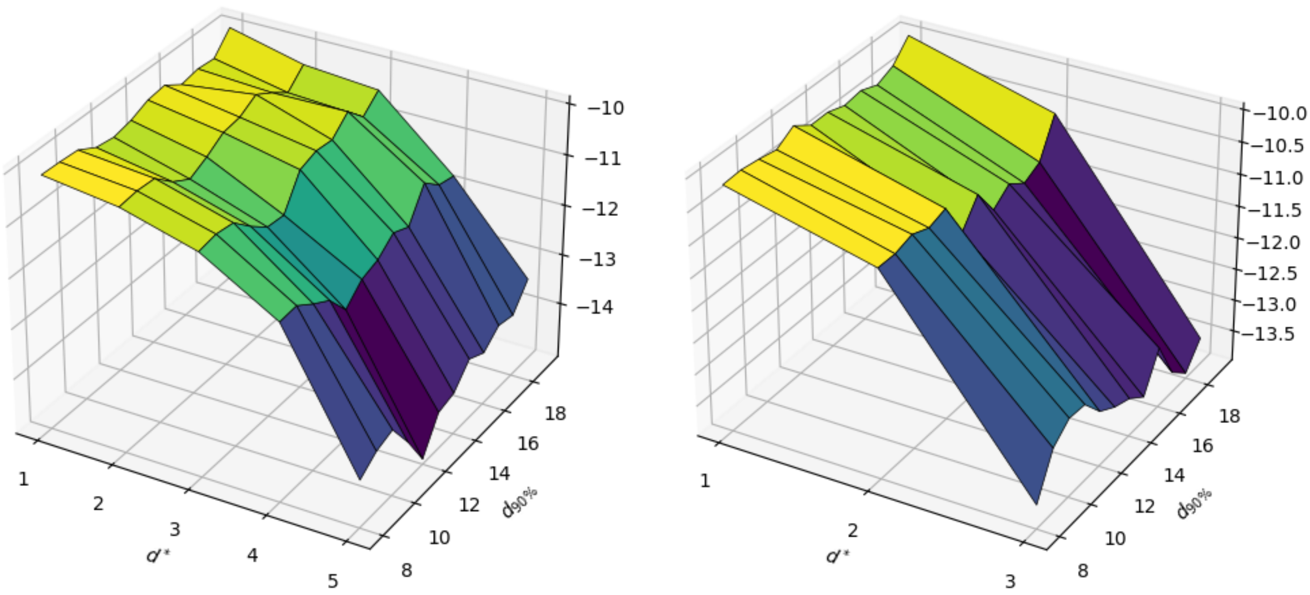
Iterating over all values of γ enables the calculation of the global minimum of the AIC. For each γ and lag of VAR given, the number of selected risk factors should be smaller than d_{L_1} or d_{L_2}, and risk factors with high average RFE result will be selected first. The optimal γ and the lag parameter of VAR models are determined by choosing the one with the lowest AIC.

The observations from Figures 10(a) and 10(b) suggest that the VAR(1) model with five risk factors corresponding to γ values in the interval $[44, 69)$ is the optimal choice.

Appendix C Justification of utilising PCA

In the subsequent simulations, we investigate various scenarios related to the prediction of house prices, examining the efficacy of PCA at the national level model when PCA is not needed. We utilise a framework comprising eight stationary time series labelled from `A` to `H`. Their interdependencies are modelled using a VAR(1) framework, the parameters of which are randomly generated. Within this construct, the variable $n_{\text{not related}}$ indicates the precise number of time series (ranging from `A` to `G`) that remain uncorrelated with series `H`. In addition, four distinct cases are presented:

- Case 1: A baseline scenario where PCA is not applied.
- Case 2: PCA is employed to series (ranging from `A` to `G`), and the model retains 99% of the variance.
- Case 3: The proportion of variance retained using PCA is adjusted to 95%.
- Case 4: The proportion of variance retained using PCA is adjusted to 90%.



(a) AIC of VAR(1) at the SA4 Level

(b) AIC of VAR(2) at the SA4 Level

Figure 10. AIC of VAR models with different lags. The values of AIC change with two parameters: the smallest number of risk factors explaining at least 90% of the variance $d_{90\%}$, and the selected number of risk factors d^*.

Table 11 presents the outcomes from 100,000 simulations. A close examination reveals a marked similarity between Case 1 and Case 2 in terms of prediction accuracy. Notably, when only a few time series remain uncorrelated to series `H`, employing PCA tends to adversely affect the model's performance. Yet, as the count of these uncorrelated series grows, the disparity in results diminishes rapidly.

Table 11. Comparison of the MSE of predicted house prices from the VAR model incorporated with PCA at the national level in the worst case.

$n_{\text{not related}}$	Case 1	Case 2	Case 3	Case 4
0	8.8096	8.8848	10.0716	10.2878
1	8.7439	8.0961	9.6170	9.7390
2	8.0615	8.2594	8.0304	9.2413
3	6.6619	6.7089	7.2845	7.2887
4	5.9108	5.8087	6.0204	6.2489
5	4.6003	4.5455	5.1078	4.7525
6	2.9982	3.1134	3.0481	3.3846
7	1.3360	1.3328	1.3386	1.3539

Notes: Each generated time series spans a length of 200, with the initial 80% designated as training data and the remaining 20% serving as the test dataset.

It is important to highlight that the described scenario represents the least favourable conditions for implementing PCA. In situations other than this extreme, the variables in the VAR(1) model are no longer time series `A` to `H`. On the contrary, the variables consist of several latent risk factors and series `H`, and the number of risk factors is assumed to be 5. The variable $n_{\text{not related}}$ represents the number of the variables within the time series from `A` to `G` that are not related to the generated latent factors. All remaining time series are assumed to be random linear combinations of these latent risk factors.

Table 12 shows the comparison of MSE in general cases. For $n_{\text{not related}} = 0$, implying all variables from `A` to `G` are related to latent factors, the MSE is lowest for Case 1 and highest for Case 4. This indicates that the raw data without PCA application may already have strong interpretability for this scenario. As $n_{\text{not related}}$ increases (from 0 to 6), the MSE for Case 1 also shows an increasing trend, implying that as more unrelated variables are introduced, the model's accuracy without PCA diminishes. In contrast, Cases 2 to 4, which involve PCA, show that the MSE does not strictly increase with the number of unrelated variables. This suggests that PCA helps mitigate the noise introduced by unrelated variables, especially when a higher variance is retained (as in Case 2). Across all values of $n_{\text{not related}}$, Case 2 generally has MSE values close to Case 1. This implies that retaining 99% variance using PCA produces results similar to the no-PCA scenario, but potentially with reduced complexity.

Table 12. Comparison of the MSE of predicted house prices from the VAR model incorporated with PCA at the national level in general cases.

$n_{\text{not related}}$	Case 1	Case 2	Case 3	Case 4
0	9.4297	9.8874	9.9220	10.6870
1	9.6192	9.9785	9.9847	10.7113
2	10.0314	10.0125	10.0985	11.2724
3	10.3159	10.0949	10.1785	11.3367
4	10.8601	10.1044	10.2565	11.6464
5	11.1843	10.8126	10.7289	11.8777
6	11.7172	10.9306	12.4313	11.9736

Notes: Each generated time series spans a length of 200, with the initial 80% designated as training data and the remaining 20% serving as the test dataset.

Appendix D FAVAR

The available number of HPI growth rates in area j at the SA4 level is collectively denoted by the $|\Omega_j| \times 1$ vector $\boldsymbol{h}_t$, which consists of house prices index in area j at the SA4 level. $|\Omega_j|$ is greater than the total number of latent factors, K, and observed variables M in the FAVAR system. The vector of the observed HPI growth rate $\boldsymbol{h}_t$ is assumed to be related to the unobservable SA4 level latent factors $\boldsymbol{f}_t^L(\Omega_j)$ and the country-level risk factors (including the national house price growth rate) $\boldsymbol{f}_t^P(\Omega)$ in the hierarchical model by an observation equation of the form:

$$\boldsymbol{h}_t = \Lambda^L(\Omega_j)\boldsymbol{f}_t^L(\Omega_j) + \Lambda^P \boldsymbol{f}_t^P(\Omega) + \xi_t, \tag{32}$$

where $\Lambda^L(\Omega_j)$ is an $|\Omega_j| \times K$ matrix of factor loadings, Λ^P is an $|\Omega_j| \times M$ matrix, and the $|\Omega_j| \times 1$ vector of error terms ξ_t are mean zero and assumed to be normal to display a small amount of cross-correlation. The country-level risk factors $\boldsymbol{f}_t^P(\Omega)$ are treated as observed variables, which illustrates that $M = 7$ and the joint dynamics of $\left(\boldsymbol{f}_t^L(\Omega_j), \boldsymbol{f}_t^P(\Omega)\right)$ are given by:

$$\begin{bmatrix} \boldsymbol{f}_t^L(\Omega_j) \\ \boldsymbol{f}_t^P(\Omega) \end{bmatrix} = \Phi(L) \begin{bmatrix} \boldsymbol{f}_{t-1}^L(\Omega_j) \\ \boldsymbol{f}_{t-1}^P(\Omega) \end{bmatrix} + \boldsymbol{v}_t, \tag{33}$$

where $\Phi(L)$ is a conformable lag polynomial of finite order d, and the error term, $\boldsymbol{v}_t$, is mean zero with a covariance matrix. A FAVAR model, denoted by FAVAR1, with $K = 6$ is built for further comparison with the hierarchical model, and the lag for this FAVAR model is selected to be 1 in the VAR order selection process.

The FAVAR2 model is constructed to measure the growth rates of all-HPI at the national level. This is represented by a vector $\boldsymbol{h}_t$ of size $|\Omega| \times 1$, where $K = 6$ and the lag is 1. The model's observation equation is given by:

$$\boldsymbol{h}_t = \Lambda^L \boldsymbol{f}_t^L(\Omega) + \Lambda^P \boldsymbol{f}_t^P(\Omega) + \xi_t. \tag{34}$$

Here, Λ^L is an $|\Omega| \times K$ matrix of factor loadings, Λ^P is an $|\Omega| \times M$ matrix, and ξ_t is a vector of error terms of size $|\Omega| \times 1$. These error terms are assumed to be mean-zero, normally distributed, and exhibit a small cross-correlation. The joint dynamics of $\left(\boldsymbol{f}_t^L(\Omega), \boldsymbol{f}_t^P(\Omega)\right)$ is expressed as:

$$\begin{bmatrix} \boldsymbol{f}_t^L(\Omega) \\ \boldsymbol{f}_t^P(\Omega) \end{bmatrix} = \Phi(L) \begin{bmatrix} \boldsymbol{f}_{t-1}^L(\Omega) \\ \boldsymbol{f}_{t-1}^P(\Omega) \end{bmatrix} + \boldsymbol{v}_t. \tag{35}$$

Appendix E Analysis of risk factors

Table 13. Comparison of suburbs with the highest and lowest values of $\beta_i^P(0)$ in Proposition 3.3.

Postcode	Longitude	Latitude	$\beta_i^P(0)(\times 10^{-3})$	Suburb name
10 suburbs with the highest $\beta_i^P(0)$				
3104	145.09	-37.78	7.62	Balwyn North, Greythorn
3108	145.12	-37.78	7.06	Doncaster
3107	145.11	-37.76	6.54	Templestowe Lower
3021	144.80	-37.75	6.20	Albanvale, St Albans, Kealba, Kings Park
3125	145.11	-37.85	6.09	Burwood, Surrey Hills South, Bennettswood
3132	145.19	-37.81	6.04	Mitcham North, Mitcham, Rangeview
3133	145.20	-37.85	5.83	Vermont, Vermont South
3066	144.99	-37.79	5.82	Collingwood, Collingwood North
3928	144.97	-38.40	5.73	Main Ridge
3127	145.09	-37.82	5.57	Surrey Hills, Surrey Hills North, Mont Albert
10 suburbs with the lowest $\beta_i^P(0)$				
6041	115.55	-31.33	-6.68	Wilbinga, Gabbadah, Caraban, Woodridge
6450	122.05	-33.93	-6.98	Chadwick, Condingup, Nulsen, Coomalbidgup
6168	115.69	-32.24	-6.98	Hillman, Rockingham, Peron, Garden Island
6230	115.66	-33.35	-7.20	Withers, Dalyellup, Gelorup, Vittoria
6280	115.30	-33.71	-7.51	Carbunup River, Bovell, Busselton, Broadwater
4744	147.93	-22.11	-7.51	Moranbah
6208	115.84	-32.66	-7.57	Ravenswood, Pinjarra, Meelon, West Pinjarra
6210	115.72	-32.55	-8.16	Mandurah, Parklands, Bouvard, Greenfields
6530	113.97	-28.75	-8.42	Tarcoola Beach, Strathalbyn, Woorree, Wandina
6722	118.82	-20.29	-8.88	Boodarie, Finucane, De Grey, Pippingarra

Table 14. Comparison of suburbs with the highest and lowest coefficients of l_{t-1} in Proposition 3.3.

Postcode	Longitude	Latitude	Coefficient($\times 10^{-3}$)	Suburb name
10 suburbs with the coefficients				
3418	141.84	-36.29	8.24	Gerang Gerung, Nhill, Kiata, Lawloit
2737	142.70	-34.64	8.14	Euston
2395	149.44	-31.54	8.10	Ropers Road, Binnaway, Weetaliba
6479	118.23	-30.77	7.93	Lake Brown, Dandanning, Barbalin, Mukinbudin
3226	144.52	-38.27	7.85	Ocean Grove
3231	144.03	-38.47	7.79	Aireys Inlet, Eastern View, Fairhaven
7300	147.22	-41.61	7.77	Perth, Devon Hills, Powranna
2800	148.99	-33.35	7.73	Lucknow, Nashdale, Summer Hill Creek, Huntley
5414	138.82	-33.96	7.71	Manoora
2026	151.27	-33.89	7.71	Tamarama, Bondi Beach, Bondi, North Bondi
10 suburbs with the lowest coefficients				
3108	145.12	-37.78	6.36	Doncaster
3678	146.41	-36.50	6.34	Oxley Flats, Rose River, Edi, Meadow Creek
3869	146.37	-38.37	6.30	Yinnar, Yinnar South, Jumbuk
5495	138.04	-32.91	6.29	Baroota, Port Germein, Mambray Creek
3289	142.19	-37.89	6.27	Gazette, Tabor, Gerrigerrup, Penshurst, Purdeet
5540	138.01	-33.19	6.25	Bungama, Pirie East, Risdon Park, Port Davis
4405	151.32	-27.12	6.24	Ducklo, Dalby, Blaxland, Bunya Mountains
3723	146.15	-37.15	6.23	Howes Creek, Howqua Inlet, Gaffneys Creek
3282	142.39	-38.31	5.95	Koroit, Illowa
5680	134.14	-32.65	5.64	Chinbingina, Sceale Bay, Streaky Bay

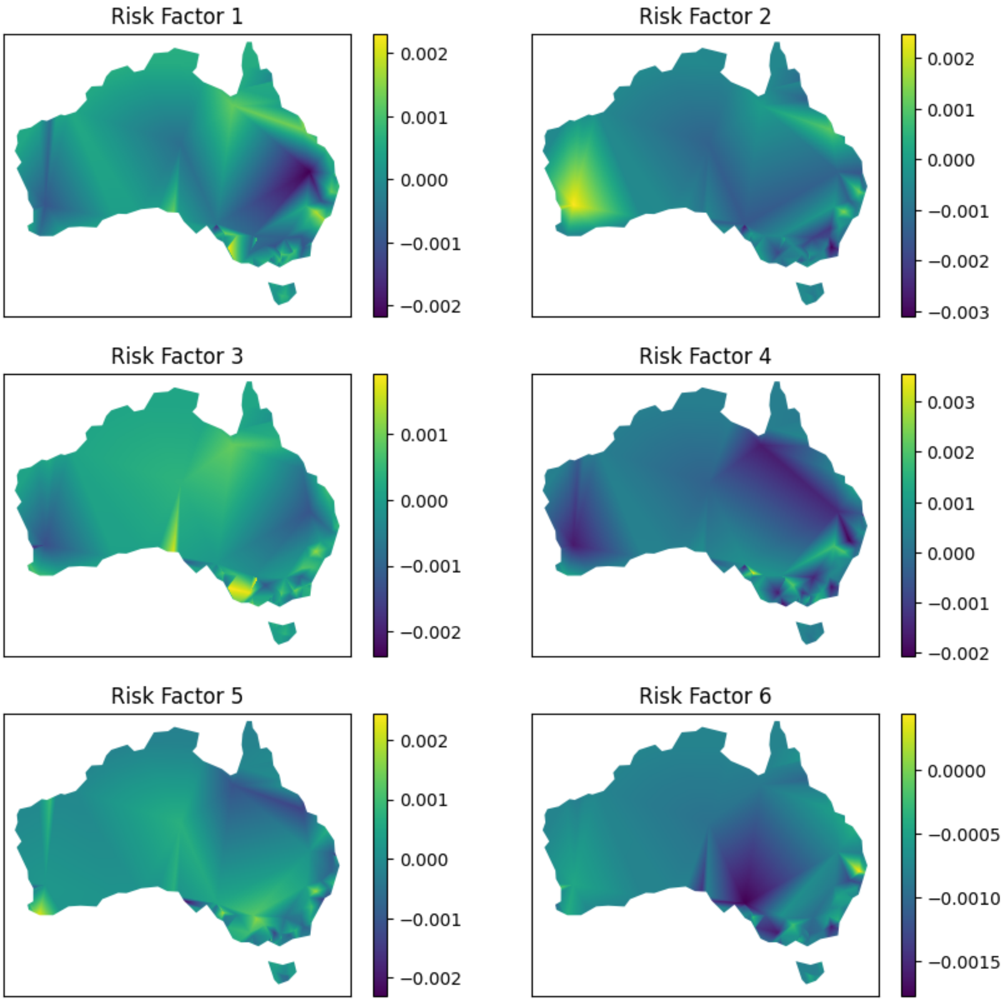


Figure 11. Coefficients of national risk factors l in the previous period in each suburb.

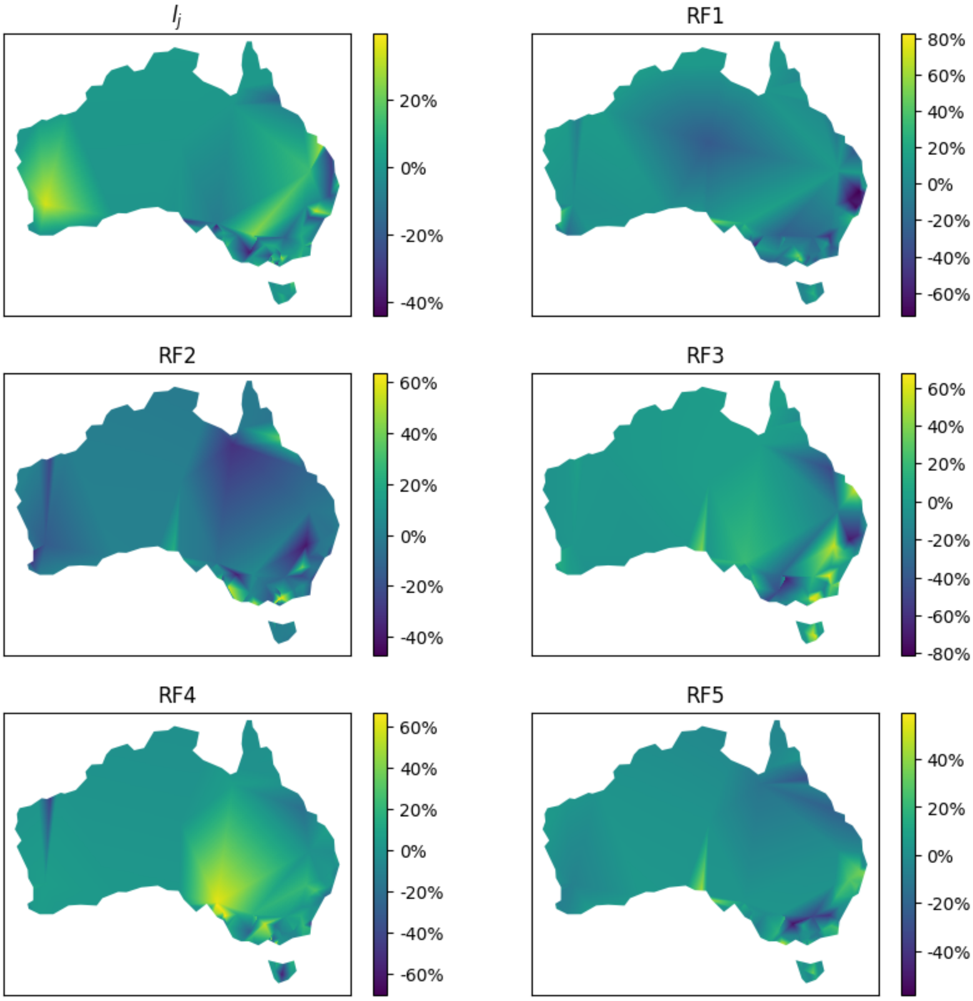
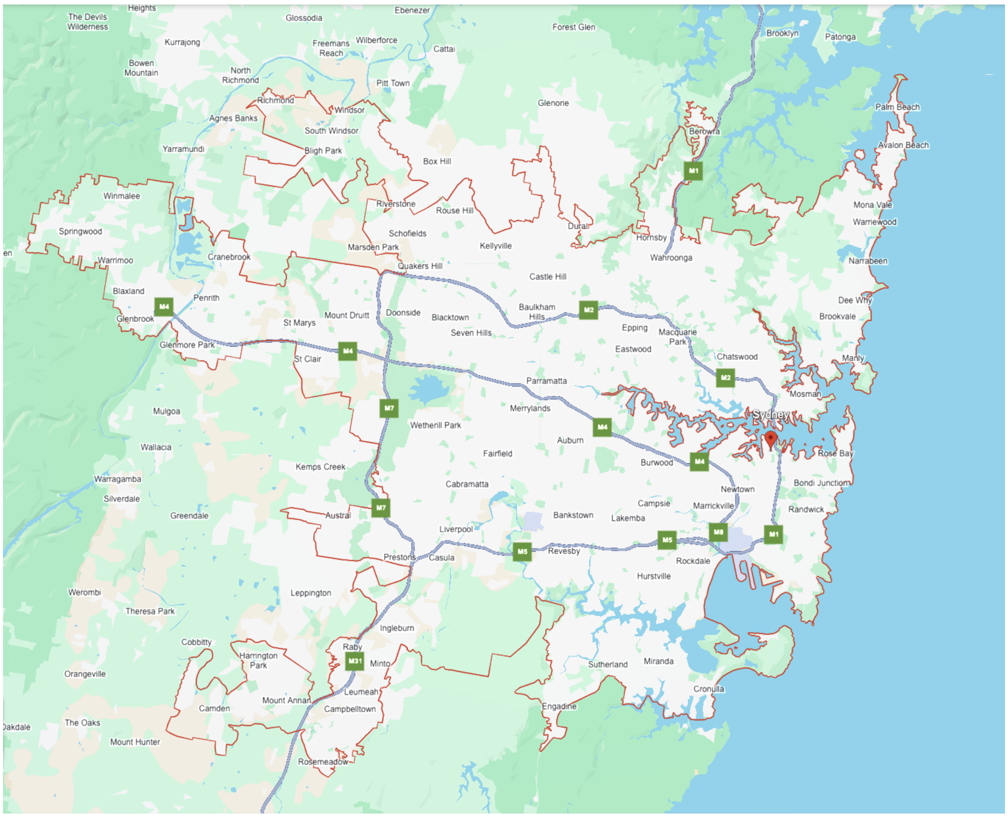
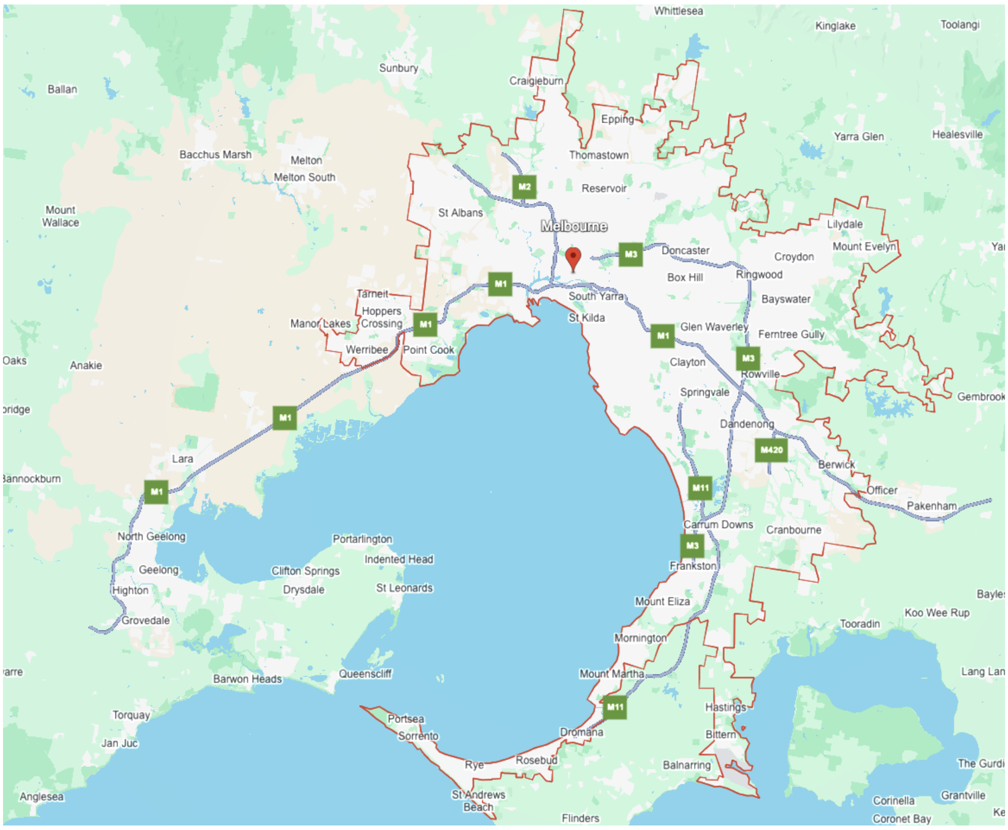


Figure 12. The average historical contribution of risk factors at the SA4 level to each suburb's overall house price changes. The average historical contribution of risk factors is defined in Equation (23).

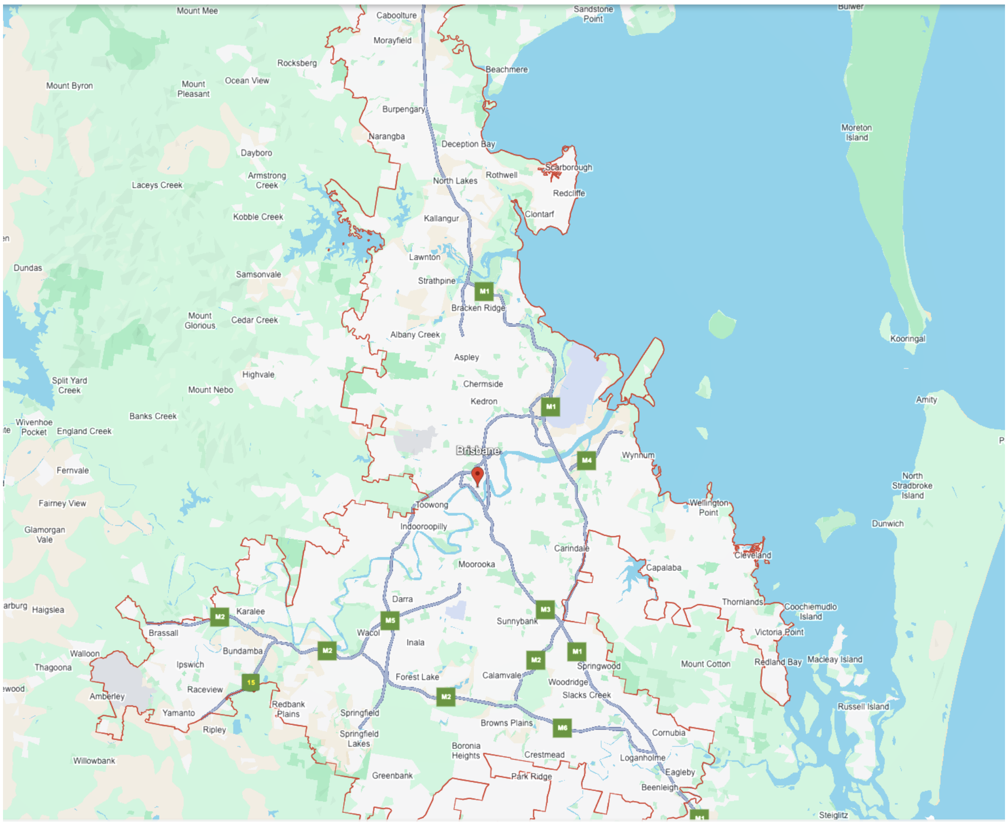
Appendix F Sydney, Melbourne, Brisbane, and Perth metropolitan areas



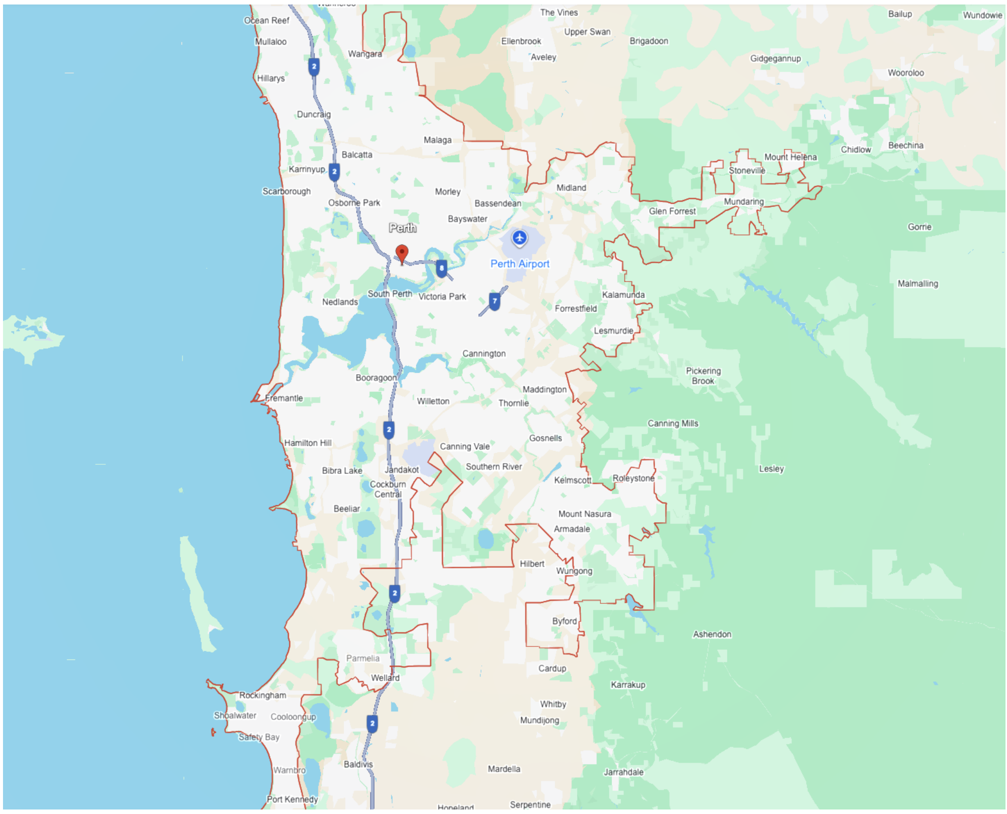
(a) Sydney



(b) Melbourne



(c) Brisbane



(d) Perth

Figure 13. Sydney, Melbourne, Brisbane, and Perth metropolitan areas. Maps sourced from Google Earth (Google, 2024).