ARC CENTRE OF EXCELLENCE IN
POPULATION AGEING RESEARCH

ARC Centre of Excellence in Population Ageing Research

Working Paper 2023/15

Hierarchical House Price Model incorporating Geographical and Macroeconomic Factors

Lingfeng Lyu

Hierarchical House Price Model Incorporating Geographical and Macroeconomic Factors

Lingfeng Lyu, Jonathan Ziveyi, Michael Sherris, Yang Shen

Abstract

This paper presents a tri-level hierarchical approach to house price modelling at the postcode level, which is considered the most granular geographical scale, incorporating macroeconomic influences from the national level and integrating data from the largest sub-state level (SA4). By employing a Risk Premium - Principal Component Analysis (RP-PCA) for SA4-level risk factors and combining these with national-level risk factors, a vector autoregressive (VAR) model is developed. This geographically conditional multi-factor model with a hierarchical structure offers enhanced short-term prediction accuracy while maintaining long-term forecasting capabilities. The model's predictive accuracy is further enhanced by introducing an empirical copula to describe the dependence structure of one-step residuals across various suburbs. This methodology grants a dynamic and granular view of housing price trends in Australia. Key determinants like interest rate shifts, GDP growth, and exchange rate variances play a crucial role, particularly in urban areas in metropolitan cities. The analysis of economic and demographic factors on the SA4 level indicates that elements such as home debt increments, wage fluctuations, and population shifts are pivotal in shaping housing prices, underscoring the significance of a granular regional analysis.

Keywords: House price modelling, Hierarchical framework, Macroeconomic variables, Risk Premium - Principal Component Analysis (RP-PCA).

1 Introduction

The literature on house price modelling and prediction is vast and diverse, encompassing various techniques and methodologies to capture the complex dynamics of housing markets. A stream of literature focuses on the role of macroeconomic variables and their impact on house prices. Goodman and Thibodeau (2008) investigate the relationship between macroeconomic factors and house price dynamics, while Kuttner and Shim (2012) analyse the role of monetary policy and credit market conditions in influencing house prices in Australia. Furthermore, Adams and Füss (2010) examine the linkages between housing and economic growth in different countries, highlighting the importance of considering both global and local factors when modelling house prices. There has also been growing interest in incorporating local risk factors into house price models. Green and Malpezzi (2001) analyse the impact of local labour market conditions on house prices. Guerrieri et al. (2013) explore the role of local credit conditions and lending practices. More recently, Glaeser and Gyourko (2018) highlight the significance of local amenities and public goods in determining house prices, stressing the importance of accounting for these factors in house price models. Although these studies investigate the risks in different regions, the introduction of these variables primarily serves as an explanation for the factors influencing house prices. The potential relationships among these variables and how to incorporate them into the model for house price prediction are not thoroughly explored in these articles.

This study presents a hierarchical framework for modelling house prices utilising a wider range of data. Prior studies have suggested that risk factors related to national-level variables such as zero-coupon bond yield rates, gross domestic product growth rates, house price growth rates, and rental yield rates exhibit temporal co-dependence(Hanewald and Sherris, 2013). One aim of the house price model is to incorporate a wider range of risk factors on the national level. The spatial autoregressive model in LeSage (2008) indicates the feasibility of using the Principal Component Analysis (PCA) to find latent risk factors that are influential in house price models. In addition, drawing inspiration from the factor-augmented vector autoregression (FAVAR) model proposed in Bernanke et al. (2005), a more pragmatic approach is proposed to formulate risk factors based on variables that cannot be directly observed like the aforementioned macroeconomic variables in this study. Micro-level risk factors are exogenously incorporated to supplant the majority of latent factors in the FAVAR model, thereby depicting the house price index for a specific area, such as a suburb. In this study, the desirability for home equities in each area is assumed to be influenced by changes in the demographics of the region, income and wealth levels, and the structure of the housing stock (Chomik and Yan, 2019). Therefore, one of the most significant contributions of this study is analysing local risk factors dynamically separately from other fixed features.

This study employs a geographically conditional CAPM model, adapted from the model in Jagannathan and Wang (1996), to develop a hierarchical framework, enabling the integration of various levels of risk factors. The inclusion of these risk factors in the model through time series analysis aids in elucidating and forecasting the average growth rate for each area, as opposed to relying solely on the national-level average. As a result, this study offers a more precise representation of excess house price growth rates across disparate areas, in line with prior economic model assumptions. Besides its capacity to incorporate an increased number of risk factors, the model's distinct structure facilitates the improved introduction of mixed-frequency data at different geographical levels. This model integrates a comprehensive set of variables, bolstering interpretation and short-term prediction accuracy while initially safeguarding long-term forecasting capabilities from significant deterioration, a critical aspect for further analysis of residence-related insurance products. To circumvent overfitting issues, this study constrains the number of parameters in the hierarchical model by employing the adjusted PCA method with penalty terms (Lettau and Pelger, 2020). Moreover, the utilisation of information from a model's residual covariance structure has been demonstrated to enhance prediction accuracy, as evidenced by studies such as Wickramasuriya et al. (2018). Therefore, to further enhance model accuracy, this project decomposes the residuals of the primary house price model into two components based on theoretical analysis. In the end, the final model consists of a linear model characterising the dynamic differences in desirability for residential property by location, while a multi-variable copula methodology is adapted to depict the model's volatility, demonstrating the conditional covariances across various areas resulting from missing information in data collection.

The remaining part of the paper is structured as follows: Firstly, the data used in the study is described and preprocessed to ensure its suitability for analysis. In the problem statement section, the research objectives and overall aim are outlined to provide a solid foundation. The hierarchical house price model is then discussed in detail, elucidating its components and the relationships between hierarchical levels. The formulation of risk factors at the national and SA4 levels is meticulously explained, highlighting their significance and the methodologies employed. Subsequently, a comprehensive volatility analysis is conducted, utilising various techniques and presenting numerical results. The numerical section concludes with the key findings and their implications. The predictive accuracy of the hierarchical model, incorporating volatility analysis, is compared to the Shrink(Mint) model, with the analysis's validity being assessed. The appropriateness of employing a hierarchical model, drawing risk factors from principal components, is also evaluated in comparison to the FAVAR model, supplemented by robustness tests for the adapted RP-PCA. Additionally, a deeper exploration of the derived risk factors is presented in the numerical section, and a sensitivity analysis of the final model is utilised to illustrate economic interpretation. In the conclusion and discussion section, the implications of the findings are thoroughly discussed, potential limitations of the study are addressed, and suggestions for future research directions are provided.

2 Data

The Australian Statistical Geography Standard (ASGS) is a social geography that categorises Australia into a hierarchy of statistical areas based on the location of people and communities. The ABS Structures are geographies specifically designed by the ABS for statistical release and analysis, taking into account the requirements of statistical collections and relevant geographic concepts. The ABS Structures include six interrelated hierarchies of regions: Mesh Blocks (MBs), Statistical Areas Level 1 (SA1s), Statistical Areas Level 2 (SA2s), Statistical Areas Level 3 (SA3s), Statistical Areas Level 4 (SA4s), States and Territories (S/T), and Australia (AUS). The ASGS was introduced in 2011 to replace the previous standard, the Australian Standard Geographical Classification (ASGC), which had been in use since 1984. Due to frequent changes in geographical classification standards in recent decades, postcodes are used as a benchmark to collect variables from suburbs as they have not undergone significant classification changes. Variables are estimated at different levels, with the postcode level being the lowest level where data can be collected.

CoreLogic is a prominent corporation that specialises in providing standard data and other related services pertaining to home equity. The company offers a monthly postcode-level home equity value index and associated rental yield indices for various regions in Australia, with data dating back to January 1980. As shown in Figure 1, these datasets have been pre-processed such that all indices are standardised to 100 in December 2009, thereby preventing direct comparison of absolute house price indices across different regions based on CoreLogic data alone. Nevertheless, the realistic median house price in each area provides a means for converting house price growth rates to realistic house prices. Consequently, the indices' changes in different areas, which can be derived, serve as targets for house price models.

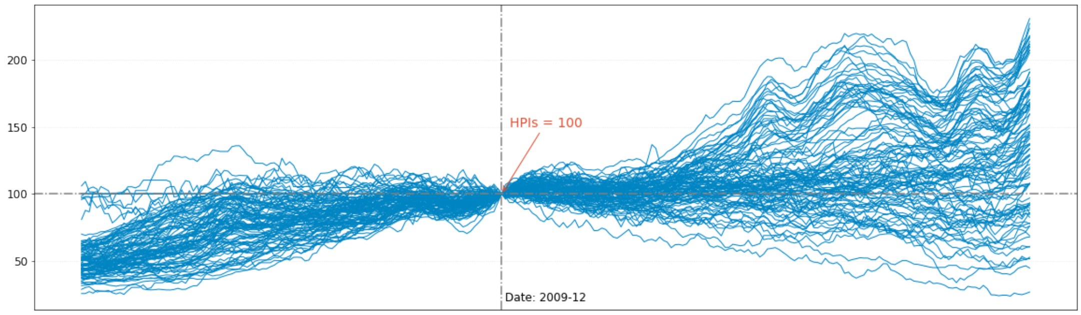


Fig. 1. The Monthly House Price Index (HPI) series in all suburbs, spanning from January 2000 to December 2020. The x-axis represents time, while the y-axis represents the HPI. The HPI data is sourced from CoreLogic.

The study first aimed to incorporate relevant risk factors that could affect the whole housing market. The Reserve Bank of Australia (RBA) was consulted, and a number of common macroeconomic variables were identified as potentially significant, including the inflation index, exchange rate, GDP index, and risk-free interest rate.

According to Chomik and Yan (2019), a range of variables can impact the price of housing by influencing the underlying supply and demand dynamics. Specifically, the authors identify four commonly acknowledged drivers of house prices. Firstly, population factors have been found to play a key role, with increases in the number of people tending to drive up demand for housing. Recent and temporary migrants, on the other hand, are more likely to rent rather than buy, which can have implications for the overall housing market (Australian Bureau of Statistics, 2022). Secondly, the deregulation of banking is identified as a significant driver of demand for housing. This has allowed people to access credit more readily in a low-inflation, low-interest environment, which can boost demand for housing and push up prices. Thirdly, the responsiveness of housing supply is also found to be a key driver of house prices. In Australia, the price responsiveness of supply tends to be relatively slow, particularly prior to the COVID-19 pandemic. Lastly, the income level of individuals is identified as a key driver of house prices, with modelling indicating that an extra dollar of income can increase house prices in New South Wales by more than a dollar. This suggests that price sensitivity is an important factor in the relationship between income level and house prices. Taken together, these four drivers highlight the complex interplay of factors that can impact the price of housing and underscore the need for careful consideration of multiple variables when attempting to understand regional housing market trends.

Therefore, a range of variables that are believed to influence the regional supply and demand for houses in different areas is collected. Specifically, annual data was collected from the Household, Income and Labour Dynamics in Australia (HILDA) Survey, which provides insights into the demographics, income levels, and employment status of residents in various districts. Population-related variables, which may include variables such as age, gender, family size, language, education level, and employment status, are identified as key drivers of house prices. Annual income is likely to be closely linked to the fourth driver of house prices identified by Chomik and Yan (2019), namely income level. Additionally, wealth and debt may be linked to the second driver of house prices, namely banking deregulation and access to credit. It is worth noting that the longitude and latitude data can be utilised to characterise the underlying demand for housing (Liu et al., 2016; Pace et al., 2000), which in turn serves as an indicator of the housing supply in a particular area to some degree. The coordinates themselves will not be incorporated directly as variables in the model. However, they will be employed in the numerical analysis section as a point of reference for conducting further investigation into the geographical variation of house prices.

The variables previously mentioned are catalogued in Table 1, accompanied by their sources. As indicated in the table, 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 subsequently.

Tab. 1. Sources of Data.

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	Quaterly	RBA
	S&P/ASX 200	Monthly	Yahoo Finance

Categorical variables will be processed as the proportion of each variable in each area at different levels. The total number of individuals in the HILDA survey is incorporated as a parameter in the model. This inclusion is motivated by the fact that the sample from each area in the HILDA data is designed to be representative of the broader Australian population. As a result, the sample size may vary across different areas depending on their characteristics and the overall population (Melbourne Institute, 2019). After testing and ruling out other sources of non-stationarity, such as unit roots or seasonality, the variables or their first/second-order differences are added to the variable list. This is a standard procedure in Econometrics for dealing with non-stationary time series (Lutkepohl and Kratzig, 2004). Given that the Household, Income and Labour Dynamics in Australia (HILDA) data is only accessible starting from the year 2000, the CoreLogic house price index data, which constitutes the dependent variable in this study, are truncated. As such, the research period is confined to the time frame between January 2000 and December 2020, when the incorporation of HILDA data is in demand. The final model was built using data from the last 15 years (2006-2020), considering some variables need to be preprocessed, and some HILDA variables chosen started from 2005.

Tab. 2. Description of Variables.

Variable Name	Description	Stationary Method
ir	Interest Rate	1storderdiff
exr	Exchange Rate	1storderdiff
cpi	Consumer Price Index	1storderdiff
gdp	Gross Domestic Product	1storderdiff
rs	Retail Sales	1storderdiff
pda	Private Dwelling Approvals	1storderdiff
asx	Australian Securities Exchange	1storderdiff
hsdebt	Total Home Debt	1storderdiff
wsce	Current weekly gross wages & salary	1storderdiff
agemedian	Median of age	original
agestd	Standard deviation of age	1storderdiff
ageq3	The third quantile of age	1storderdiff
agecount	Total number of people	1storderdiff
xphmrna	Home repairs/renovations/maintenance	1storderdiff
xphltpa	Fees paid to health practitioners	2ndorderdiff
lsemp	Weekly time on paid employment	1storderdiff
iprbwm	Went without meals	original
fiprbmr	Could not pay the mortgage or rent on time	1storderdiff
ancob_1101	Country of Birth: Australia	original
ancob_1201	Country of Birth: New Zealand	1storderdiff
ancob_7103	Country of Birth: India	1storderdiff
ancob_6101	Country of Birth: China (excludes SARs)	1storderdiff
ancob_5105	Country of Birth: Vietnam	2ndorderdiff
edhigh1_4	Highest education level achieved: Year 12 or equivalent	original
edhigh1_5	Highest education level achieved: Cert III/IV	1storderdiff
edhigh1_8	Highest education level achieved: Bachelor degree	1storderdiff
edhigh1_9	Highest education level achieved: Postgraduate degree	1storderdiff
anatsi_2	Aboriginal	2ndorderdiff
anatsi_3	Torres Strait Islander	1storderdiff
chkb12_1	Employment status	2ndorderdiff

[†] The information presented in the first and second sections pertains to numerical variables, whereas the information in the last sections pertains to categorical variables. For categorical variables, the data is processed to show the proportion of each group present in a given area.

The decision to limit the number of levels to three was motivated by the desire to keep the model easily interpretable and to avoid increasing model complexity, which can negatively impact forecast accuracy. To ensure that the HILDA data corresponds to the CoreLogic data for each postcode, the middle-level data must be identified. Based on the fact that CoreLogic contains more postcodes than HILDA, a requirement for selecting the middle level is proposed: for each postcode area (in CoreLogic), the corresponding middle-level area must be found in HILDA. This means that even if the same postcode is not included in HILDA, other postcodes belonging to the same middle level must exist so that the risk factors for this middle level can be taken into account based on the data from other postcode areas. The SA4 level was found to be the most granular level that meets this requirement, with only one SA4-level area `White-Bay` being excluded. The annual SA4-level data will first be treated as a whole and then extrapolated to monthly frequency before being incorporated into the full model.

3 A dynamic house price model framework with hierarchical structure and volatility analysis

The house price index in Australia is viewed as being nested within different geographical jurisdictions hierarchically. Our basic model setup consists of two parts. The first part is a three-level hierarchical model describing house price index growth rates, where an adapted Risk-Premium-PCA (RP-PCA) model is applied to find the risk factors at the SA4 level(Lettau and Pelger, 2020). Then, an empirical copula is incorporated into the model to analyse residuals from the hierarchical model.

3.1 Problem statement

A hierarchical factor model is designed to describe the factors representing the underlying idiosyncratic desirability for land, which follow time series processes. Data at different levels are collected to estimate these factors. A three-level hierarchical model based on national, SA4-level, and postcode-level data is proposed in this study.

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

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

Here Ω is the universal set with all suburbs in Australia, Ω_j includes all suburbs belonging to state j, and Ω_{ji} is a set that contains one element, which is the i^{th} belonging to state j. The hierarchical *model keeps adding information at lower levels until collecting any exogenous information becomes impossible.*

Postcode-level excess house price index growth rates are assumed to be equivalent to excess asset returns, described by the model proposed in Fama and MacBeth (1973), estimating risk premia in the multi-factor Capital Asset Pricing Model (CAPM) setting. Static CAPM models sometimes fail to estimate the influence of underlying risk accurately because risk factors are assumed to be constant over different time periods and areas. Such an assumption is too restrictive for a house price model. Therefore, the assumption that the CAPM holds in a conditional sense in the analysis of house price index growth rates is made.

Lemma 3.1. *For the excess house price index growth rate of suburb i in period t,*

$$\mathbb{E}_\omega\left[h_{i,t} \mid I_{i,t-1}\right] = \mathbb{E}_\omega\left[\mathbf{w}_i \mid I_{i,t-1}\right]^\top \mathbb{E}_\omega\left[\boldsymbol{f}_{i,t} \mid I_{i,t-1}\right], \tag{2}$$

where $I_{i,t-1}$ is defined as the complete information set of the i^{th} suburb at the end of period $t-1$, $\mathbf{w}_i$ is a vector of coefficients describing risk premia, and $\boldsymbol{f}_{i,t}$ is a vector of risk factors for the i^{th} suburb.

Proof. Jagannathan and Wang (1996) assume that CAPM holds in a conditional sense, that is, betas and the market risk premium vary over time, which is described by Equation (2) in that article. The variation of the house price model across regions rather than across time intervals is focused. Therefore, adjustments are made to their model by assuming a spatially conditional CAPM model; the price of risks and risk factors vary across areas. □

In order to explain the cross-regional variations in the unconditional expected excess HPI growth rate in different suburbs, the unconditional expectation of both sides of Equation (2) with regard to $I_{i,t-1}$ is taken.

Lemma 3.2. *Under the assumption that each complete information sets $I_{i,t-1}$ at time t are available, the unconditional expected excess HPI growth rate is:*

$$\mathbb{E}_\omega\left[h_{i,t}\right] = \mathbb{E}_\omega\left[\mathbf{w}_i\right]^\top \mathbb{E}_\omega\left[\boldsymbol{f}_{i,t}\right] + \mathbb{C}\mathrm{ov}_\omega(\mathbb{E}_\omega\left[\mathbf{w}_i \mid I_{i,t-1}\right], \mathbb{E}_\omega\left[\boldsymbol{f}_{i,t} \mid I_{i,t-1}\right]). \tag{3}$$

Because $I_{i,t-1}$ is defined as the complete information set of the i^{th} suburb, which means $\mathbb{E}_\omega\left[\mathbf{w}_i \mid I_{i,t-1}\right] = \mathbb{E}_\omega\left[\mathbf{w}_i\right]$, and $\mathbb{E}_\omega\left[\boldsymbol{f}_{i,t} \mid I_{i,t-1}\right] = \mathbb{E}_\omega\left[\boldsymbol{f}_{i,t}\right]$. Therefore,

$$\mathbb{E}_\omega\left[h_{i,t}\right] = \mathbb{E}_\omega\left[\mathbf{w}_i\right]^\top \mathbb{E}_\omega\left[\boldsymbol{f}_{i,t}\right], \tag{4}$$

which is a classic Multi-Factor Model at time t.

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

Remark 3.1. *Assuming $\mathbb{E}_\omega\left[\boldsymbol{f}_{i,t}\right] = \mathbb{E}_\omega\left[\boldsymbol{f}_t\right]$, which means that all risk factors remain the same across different areas, the unconditional expected excess HPI growth rate becomes:*

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

which is subject to the expression for the multi-factor CAPM model. However, as aforementioned at the beginning of this section, the assumption that all risk factors remain the same is inappropriate. As the property cannot be transferred or transacted within the desired period, this low liquidity nature of real estate leads to a less strict assumption: risk factors for house prices are partially different among different areas. For example, some risk factors, such as the macroeconomic ones, remain the same across different suburbs, while some other risk factors are not. Finding these risk factors is the purpose of the hierarchical house price model.

However, the largest information set of each suburb i is incomplete, which is written as a function of the a series of hierarchical partitions: $I_{i,t-1}(\Omega, \Omega_j, \Omega_{ji}) \subset I_{i,t-1}$. The lack of information will inevitably lead to model inaccuracy.

Remark 3.2. *The definition of partitions in this context diverges from the general definitions in measure theory:*

$$I_{i,t-1}(\Omega_{ji}) \not\subset I_{i,t-1}(\Omega_j) \not\subset I_{i,t-1}(\Omega). \tag{6}$$

The inclusion relationship between different partitions is assumed to be

$$I_{i,t-1}(\Omega_{ji}) \subset I_{i,t-1}(\Omega_j, \Omega_{ji}) \subset I_{i,t-1}(\Omega, \Omega_j, \Omega_{ji}) \subset I_{i,t-1}, \tag{7}$$

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.

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:*

$$\mathbb{E}_\omega\left[h^P_{i,t} \mid I_{i,t-1}(\Omega, \Omega_j, \Omega_{ji})\right] = \mathbb{E}_\omega\left[\mathbf{w}^P_i \mid I_{i,t-1}(\Omega, \Omega_j, \Omega_{ji})\right]^\top \mathbb{E}_\omega\left[\boldsymbol{f}^P_{i,t} \mid I_{i,t-1}(\Omega, \Omega_j, \Omega_{ji})\right], \tag{8}$$

where $h_{i,t}^P$ is the partial excess house price index growth rate of Suburb i in time t. The terms $\mathbf{w_i}^P$ and $\boldsymbol{f}_{i,t}^P$ denote the partial risk premia and factors.

Subsequently, the definition of $\mathbb{E}_\omega\left[h_{i,t}\right]$ in Lemma 3.2 has evolved:

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

The convariance term cannot be omitted because risk premia and risk factors depend on $I_{i,t-1}(\Omega,\Omega_j,\Omega_{ji})$.

According to the structure of data given, decomposition of the house price model hierarchically by taking conditional expectations of Equation (8) 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:*

$$\begin{aligned}
&\mathbb{E}_\omega\left[h_{i,t}^P \mid I_{i,t-1}\left(\Omega_j,\Omega_{ji}\right)\right] \\
&= \mathbf{w}_i^P([1:k])^\top \boldsymbol{f}_{i,t}^P([1:k]) + \mathbb{E}_\omega\left[\mathbf{w}_i^P(-[1:k]) \mid I_{i,t-1}\left(\Omega_j,\Omega_{ji}\right)\right]^\top \mathbb{E}_\omega\left[\boldsymbol{f}_{i,t}^P(-[1:k]) \mid I_{i,t-1}\left(\Omega_j,\Omega_{ji}\right)\right] \\
&+ \mathbb{C}\mathrm{ov}_\omega\left(\mathbb{E}_\omega\left[\mathbf{w}_i^P(-[1:k]) \mid I_{i,t-1}\left(\Omega_j,\Omega_{ji}\right))\right], \mathbb{E}_\omega\left[\boldsymbol{f}_{i,t}^P(-[1:k]) \mid I_{i,t-1}\left(\Omega_j,\Omega_{ji}\right)\right]\right),
\end{aligned} \tag{10}$$

where $\mathbf{w}_i^P([1:k])$ and $\boldsymbol{f}_{i,t}^P([1:k])$ are the first k risk factors and premia, and $\mathbf{w}_i^P(-[1:k])$ and $\boldsymbol{f}_{i,t}^P(-[1:k])$ are the factors and premia excluding the first k ones.

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

$$\begin{aligned}
\mathbb{E}_\omega\left[h_{i,t}^P \mid I_{i,t-1}\left(\Omega_j,\Omega_{ji}\right)\right] &= \mathbb{E}_\omega\left[\mathbb{E}_\omega\left[h_{i,t}^P \mid I_{i,t-1}\left(\Omega,\Omega_j,\Omega_{ji}\right)\right] \mid I_{i,t-1}\left(\Omega_j,\Omega_{ji}\right)\right] \\
&= \mathbb{E}_\omega\left[\mathbf{w}_i^P \mid I_{i,t-1}\left(\Omega_j,\Omega_{ji}\right)\right]^\top \mathbb{E}_\omega\left[\boldsymbol{f}_{i,t}^P \mid I_{i,t-1}\left(\Omega_j,\Omega_{ji}\right)\right] \\
&\quad + \mathbb{C}\mathrm{ov}_\omega\left(\mathbb{E}_\omega\left[\mathbf{w}_i^P \mid I_{i,t-1}\left(\Omega,\Omega_j,\Omega_{ji}\right)\right], \mathbb{E}_\omega\left[\boldsymbol{f}_{i,t}^P \mid I_{i,t-1}\left(\Omega,\Omega_j,\Omega_{ji}\right)\right] \mid I_{i,t-1}\left(\Omega_j,\Omega_{ji}\right)\right),
\end{aligned} \tag{11}$$

and

$$\begin{aligned}
&\mathbb{E}_\omega\left[\mathbf{w}_i^P \mid I_{i,t-1}\left(\Omega_j,\Omega_{ji}\right)\right]^\top \mathbb{E}_\omega\left[\boldsymbol{f}_{i,t}^P \mid I_{i,t-1}\left(\Omega_j,\Omega_{ji}\right)\right] \\
&= \mathbf{w}_i^P([1:k])^\top \boldsymbol{f}_{i,t}^P([1:k]) + \mathbb{E}_\omega\left[\mathbf{w}_i^P(-[1:k]) \mid I_{i,t-1}\left(\Omega_j,\Omega_{ji}\right)\right]^\top \mathbb{E}_\omega\left[\boldsymbol{f}_{i,t}^P(-[1:k]) \mid I_{i,t-1}\left(\Omega_j,\Omega_{ji}\right)\right].
\end{aligned} \tag{12}$$

□

3.2 Hierarchical house price model

According to Proposition 3.1, if previous risk factors are considered to be influential in house prices, the basic house price index model on the national level is

$$\text{HPI of Suburb } i : h_{i,t}^{P}(\Omega) = \mathbf{w}_i^P(\Omega)^\top \boldsymbol{f}_{i,t}^P(\Omega) \sum_{\kappa>0} L^\kappa + \eta_t, \tag{13}$$

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 house price index at the national level is

$$\begin{aligned} h_{i,t} &= h_{i,t}^P(\Omega) + e_{i,t}(\Omega) + \psi_t \\ &= \bar{\mathbf{w}}_i^P(\Omega)^\top \boldsymbol{f}_t^P(\Omega) \sum_{\kappa>0} L^\kappa + \left(\mathbf{w}_i^P(\Omega)^\top - \bar{\mathbf{w}}_i^P(\Omega)^\top\right) \boldsymbol{f}_t^P(\Omega) \sum_{\kappa>0} L^\kappa + \eta_t + e_{i,t}(\Omega) + \psi_t \\ &= \bar{\mathbf{w}}_i^P(\Omega)^\top \boldsymbol{f}_t^P(\Omega) \sum_{\kappa>0} L^\kappa + \boldsymbol{\beta}_i^P(\Omega)^\top \boldsymbol{f}_t^P(\Omega) \sum_{\kappa>0} L^\kappa + \eta_t + e_{i,t}(\Omega) + \psi_t, \end{aligned} \tag{14}$$

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 models at smaller scales will continue to analyse. 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:

$$
\begin{aligned}
&\text{HPI of Suburb } i \in \Omega_j : h^P_{i,t}(\Omega_j) = l_{0,t} + l_{j,t} + \boldsymbol{\beta}^P_i(\Omega)^\top \boldsymbol{f}^P_t(\Omega) \sum_{\kappa>0} L^\kappa + \boldsymbol{\beta}^P_i(\Omega_j)^\top \boldsymbol{f}^P_t(\Omega_j) \sum_{\kappa>0} L^\kappa, \\
&\text{Average HPI: } l_{0,t} = \bar{\mathbf{w}}^P_i(\Omega)^\top \boldsymbol{f}^P_t(\Omega) \sum_{\kappa>0} L^\kappa + \eta_t, \\
&D(\Omega, \Omega_j) : l_{j,t} = \bar{\mathbf{w}}^P_i(\Omega_j)^\top \boldsymbol{f}^P_t(\Omega_j) \sum_{\kappa>0} L^\kappa + \eta_{j,t}.
\end{aligned}
\tag{15}
$$

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 $\boldsymbol{f}^P_{i,t}(\Omega_j)$ denote SA4-level risk factors. The relationship between the real house price growth index and the prediction from the two-level model is:

$$
h_{i,t} = h^P_{i,t}(\Omega_j) + e_{i,t}(\Omega_j) + \psi_t + \psi_{j,t}, \tag{16}
$$

where $\psi_{j,t}$ is the error term arising from the formulation of $\eta_{j,t}$. The relationship above illustrates the relationship between $e_{i,t}(\Omega)$ and $e_{i,t}(\Omega_j)$:

$$
e_{i,t}(\Omega) = \mathbf{w}^P_i(\Omega_j)^\top \boldsymbol{f}^P_t(\Omega_j) \sum_{\kappa>0} L^\kappa + \psi_{j,t} + e_{i,t}(\Omega_j). \tag{17}
$$

Equation (17) verifies the basic idea of this hierarchical model: collecting information at smaller scales to analyse residuals. Therefore, if more information is collected at the most granular scale, a three-level model can be accomplished. However, considering the small amount of data in each postcode-level area, the error in estimating a large number of risk factors is large. Therefore, only the national and SA4 levels are considered to build a two-level hierarchical model before the analysis of residuals. Another interpretation is the two-level hierarchical model is equivalent to a three-level model with $D(\Omega_j, \Omega_{ji}) = 0$ because the lack of ways to obtain $\boldsymbol{f}^P_{i,t}(\Omega_{ji})$ as an estimate of the corresponding risk factors in area i.

3.3 Formulation of risk factors

Risk factors generated from collected exogenous variables, along with $l_{0,t}$ $(l_{1,t})$, are assumed to follow Vector Autoregressive Moving Average(VARMA) under the assumption that previous risk factors are influential on house prices is

$$
\begin{aligned}
\boldsymbol{F}^P_{i,t}(\Omega) &= \Psi^{AR}_1(\mathrm{L}) \boldsymbol{F}^P_{i,t}(\Omega) + \Psi^{MA}_1(\mathrm{L}) \boldsymbol{\epsilon}_t, \\
\boldsymbol{F}^P_{i,t}(\Omega_j) &= \Psi^{AR}_2(\mathrm{L})_j \boldsymbol{F}^P_{i,t}(\Omega_j) + \Psi^{MA}_2(\mathrm{L})_j \boldsymbol{\epsilon}_{j,t},
\end{aligned}
\tag{18}
$$

where L > 1 is the lag operator, $\Psi_m^{AR}(\mathrm{L})$ and $\Psi_m^{MA}(\mathrm{L})$ are lag polynomial matrices at the national level, and $\Psi_m^{AR}(\mathrm{L})_l$ and $\Psi_m^{MA}(\mathrm{L})_j$ are lag polynomial matrices for area j at the SA4 level. $\boldsymbol{F}^P$ is a vector of extended risk factors, consisting of the original risk factor vector $\boldsymbol{f}^P$ and $l_{0,t}$ $(l_{1,t})$. A moving average process is used to describe the random residual series η, where $\boldsymbol{\epsilon_t}$ represents a Gaussian random variable at time t. While this assumption is not required for parameter estimates to be consistent or asymptotically normal, results are generally more reliable in finite samples when residuals are Gaussian white noise (Lütkepohl, 2005). As can be observed from the structure of the VARMA model above, the assumption is made that the risk factors on the larger level do not affect the risk factors on the granular level.

This hierarchical multi-factor model also aims to investigate the optimal strategies for choosing risk factors. Principal component analysis (PCA) will be utilised to estimate risk factors at different levels, illustrating that this multi-factor model includes more exogenous information to analyse prediction residuals obtained from the upper-level model. Since those variables collected are assumed to be stochastic processes and considered strongly correlated, PCA is an example of multivariate time series analysis. The central idea of principal component analysis is to reduce the dimension of the problem based on its statistical structure. The reduction is achieved by transforming collected variables into new uncorrelated variables, which are the risk factors in the previous hierarchical house price model.

3.3.1 Risk factors at the national level

After finding the principal components as risk factors using PCA at the national level, the average HPI growth rate can be considered another risk factor. A Vector Autoregressive Moving Average (VARMA) model is a combination of VAR and VMA models by considering both lag order and order of moving average in the model, which describes the relationship between the risk factors. The optimal parameters of the VARMA model are selected based on three commonly used information criteria: Akaike's information criterion (AIC), the Schwarz–Bayesian information criterion (BIC), and the Hannan-Quinn information criterion (HQIC). The PCA loading matrix at the national level can be found in Appendix A, where we also provide a rationale for the appropriateness of using PCA in this context.

These three information criteria values are presented in Table 3. Although AIC suggests a VARMA(2,1) model, BIC and HQIC suggest a simple VAR(1) model. Therefore, VAR(1) should be chosen as a better model for parsimony at the national level. The estimation results for the parameters and the Cholesky decomposition of the error covariance matrix in the VAR(1) model are given in Table 4.

Tab. 3. Information criteria for VARMA models with different lags at the national level.

	AIC	BIC	HQIC		AIC	BIC	HQIC
VMA(1)	3987.925	4276.723	4104.420	VARMA(1,2)	3748.643	4374.373	4001.050
VMA(2)	3978.774	4436.038	4163.225	VAR(2)	3751.895	4209.160	3936.346
VAR(1)	3774.629	4063.428	3891.125	VARMA(2,1)	3720.842	4346.573	3973.249
VARMA(1,1)	3741.851	4199.115	3926.302	VARMA(2,2)	3752.779	4546.976	4073.142

Tab. 4. Parameter estimates and covariance matrix in the VAR(1) model at national level.

Variable	PC1	PC2	PC3	PC4	PC5	PC6	l_0
			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	-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							0.321

Based on initial values of the six risk factors and l_0, parameter estimates, and the Cholesky decomposition of the error covariance matrix of the VAR(1) model, simulation for risk factors and l_0, for the period 2010-2020 are compared with the historical values of the state variables in Figure 2. The probability density function (pdf) of the simulated risk factors tends to have a similar empirical distribution of the historical data based on the VAR(1) model. One possible explanation for the minor shift of the simulated results is the standard Gaussian residuals assumption of the VARMA model, which is described by ψ_t in Equation (14).

3.3.2 Risk factors at granular levels

To account for potential variations in the construction of risk factors, Lettau and Pelger (2020) propose a method called Risk-Premium-PCA (RP-PCA) for estimating factors in high-dimensional data using statistical factor analysis. RP-PCA is a modified version of PCA that includes a penalty

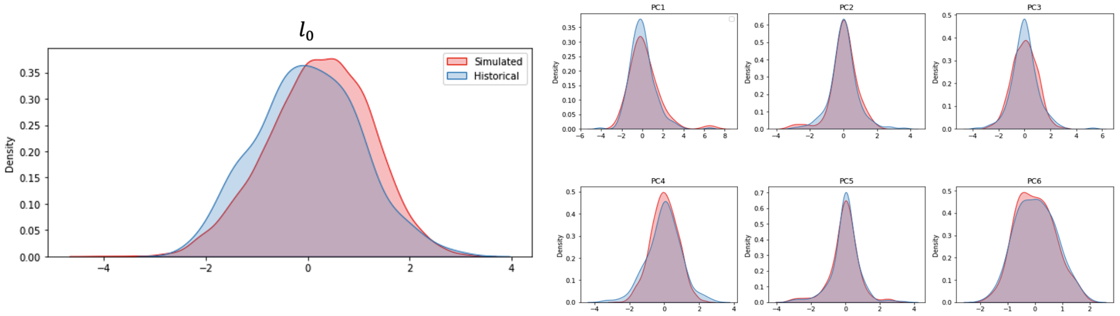


Fig. 2. Simulation results of house price index growth rates and risk factors at the national level

term to account for pricing errors. Specifically in this paper, the adapted RP-PCA penalty function includes two terms that capture the overall cross-time and cross-area errors of the PCA method. The RP-PCA model is applied to each SA4 area (Ω_j) to extract risk factors from the collected variables $X_t(\Omega_j)$:

$$\begin{aligned}(\widehat{\boldsymbol{f}}^P(\Omega_j), \widehat{\Lambda}) = \underset{\Lambda, \boldsymbol{f}^P(\Omega_j)}{\operatorname{argmin}} & \underbrace{\frac{1}{n(\Omega_j \in \Omega)T} \sum_{\Omega_j \in \Omega} \sum_{t=1}^{T} |X_t(\Omega_j) - \boldsymbol{f}_t^P(\Omega_j)\Lambda^\top|^2}_{\text{unexplained variation}} \\ & + \underbrace{\gamma_1 \frac{1}{T} \sum_{t=1}^{T} |\overline{\mathrm{X}}_t - \overline{\boldsymbol{f}}_t^P \Lambda^\top|^2}_{\text{cross-area error}} + \underbrace{\gamma_2 \frac{1}{n(\Omega_j \in \Omega)} \sum_{\Omega_j \in \Omega} |\overline{\mathrm{X}}(\Omega_j) - \overline{\boldsymbol{f}}^P(\Omega_j)\Lambda^\top|^2}_{\text{cross-time error}}.\end{aligned} \tag{19}$$

The two penalty terms in the Risk-Premium-PCA (RP-PCA) method are denoted by γ_1 and γ_2, while Λ represents the estimated PCA loadings for the middle-level model. The terms $\overline{\mathrm{X}}_t$ and $\overline{\boldsymbol{f}}_t^P$ refer to the average variables and risk factors across areas at time t, respectively. On the other hand, $\overline{\mathrm{X}}(\Omega_j)$ 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 Risk-Premium-PCA (RP-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 zero, and the objective function expressed in Equation (19) can be simplified as follows:

$$(\widehat{\boldsymbol{f}}^P(\Omega_j), \widehat{\Lambda}) = \underset{\Lambda, \boldsymbol{f}^P(\boldsymbol{\Omega_j})}{\operatorname{argmin}} \underbrace{\frac{1}{n(\Omega_j \in \Omega)T} \sum_{\Omega_j \in \Omega} \sum_{t=1}^{T} |X_t(\Omega_j) - \boldsymbol{f}_t^P(\Omega_j)\Lambda^\top|^2}_{\text{unexplained variation}} + \underbrace{\gamma_1 \frac{1}{T} \sum_{t=1}^{T} |\overline{\mathrm{X}}_t - \overline{\boldsymbol{f}}_t^P \Lambda^\top|^2}_{\text{cross-area error}}. \tag{20}$$

When the weight of the second penalty term γ_2 is set to zero, the tuning parameter γ_1 becomes the only parameter that needs to be adjusted. This parameter can be used to either underweight or overweight the means across different areas.

1. As the tuning parameter γ_1 approaches infinity, the means across different areas are overweighed, and the objective function given in Equation (20) can be simplified to:

$$(\widehat{\boldsymbol{f}}^P(\Omega_j), \widehat{\Lambda}) = \underset{\Lambda, \boldsymbol{f^P(\Omega_j)}}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} |\overline{\mathrm{X}}_t - \overline{\boldsymbol{f}}_t^P \Lambda^\top|^2. \tag{21}$$

It can be shown that minimising Equation (21) 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 also been demeaned. The motivation for overweighting the means across different areas at each time point is based on the assumption that the risk premium associated with a particular risk factor remains consistent across different areas.

2. When the tuning parameter γ_1 is set to zero, the objective function given in Equation (20) can be rewritten as:

$$(\widehat{\boldsymbol{f}}^P(\Omega_j), \widehat{\Lambda}) = \underset{\Lambda, \boldsymbol{f^P(\Omega_j)}}{\operatorname{argmin}} \frac{1}{n(\Omega_j \in \Omega)T} \sum_{\Omega_j \in \Omega} \sum_{t=1}^{T} |X_t(\Omega_j) - \boldsymbol{f}_t^P(\Omega_j)\Lambda^\top|^2. \tag{22}$$

In this case, the means across different areas are underweighted, indicating that the across-area errors are no longer taken into account when attempting to identify the principal components.

3. When the tuning parameter γ_1 is set to -1, the objective function given in Equation (20) can be rewritten as:

$$(\widehat{\boldsymbol{f}}^P(\Omega_j), \widehat{\Lambda}) = \underset{\Lambda, \boldsymbol{f}^P(\Omega_j)}{\operatorname{argmin}} \frac{1}{n(\Omega_j \in \Omega)T} \sum_{\Omega_j \in \Omega} \sum_{t=1}^{T} \left(\left(X_t(\Omega_j) - \overline{\mathrm{X}}_t\right) - \left(\boldsymbol{f}_t^P(\Omega_j) - \overline{\boldsymbol{f}}_t^P\right)\Lambda^\top\right)^2. \tag{23}$$

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 reason why the tuning parameter γ_1 is considered to be greater than zero in the subsequent analysis is twofold. First, as mentioned earlier, all variables have been pre-processed to ensure stationarity prior to the analysis. Second, setting γ_1 to a positive value is more convenient for interpretation purposes in the subsequent analysis.

To simplify the computation, the final objective function is defined as follows:

$$(\widehat{\boldsymbol{f}}^P(\Omega_j), \widehat{\Lambda}) = \underset{\Lambda, \boldsymbol{f}^P(\Omega_j)}{\operatorname{argmin}} \frac{1}{n(\Omega_j \in \Omega)T} \sum_{\Omega_j \in \Omega} \sum_{t=1}^{T} \left(\left(X_t(\Omega_j) + \gamma\overline{\mathrm{X}}_t\right) - \left(\boldsymbol{f}_t^P(\Omega_j) + \gamma\overline{\boldsymbol{f}}_t^P\right)\Lambda^\top\right)^2, \tag{24}$$

which is equivalent to Equation (20) when the tuning parameter γ (γ_1) is set to -1, 0, and infinity.

In fact, Equation (24) can be rewritten as:

$$\underset{\Lambda,\boldsymbol{f}^{P}(\boldsymbol{\Omega}_j)}{\operatorname{argmin}} \underbrace{\frac{1}{n(\Omega_j \in \Omega)T} \sum_{\Omega_j \in \Omega} \sum_{t=1}^{T} |X_t(\Omega_j) - \boldsymbol{f}_t^{P}(\Omega_j)\Lambda^{\top}|^2}_{\text{unexplained variation}} + \underbrace{\frac{\gamma(\gamma+2)}{T} \sum_{t=1}^{T} |\overline{X}_t - \overline{\boldsymbol{f}}_t^{P}\Lambda^{\top}|^2}_{\text{cross-area error}}, \tag{25}$$

which is mostly equivalent to Equation (20). This is because, when γ is greater than or equal to zero, the product $\gamma(\gamma+2)$ increases monotonically from zero to infinity. Therefore, adjusting the tuning parameter γ in the range [0,+∞] for Equation (24) is essentially equivalent to adjusting the tuning parameter γ_1 in the range [0,+∞] for Equation (20).

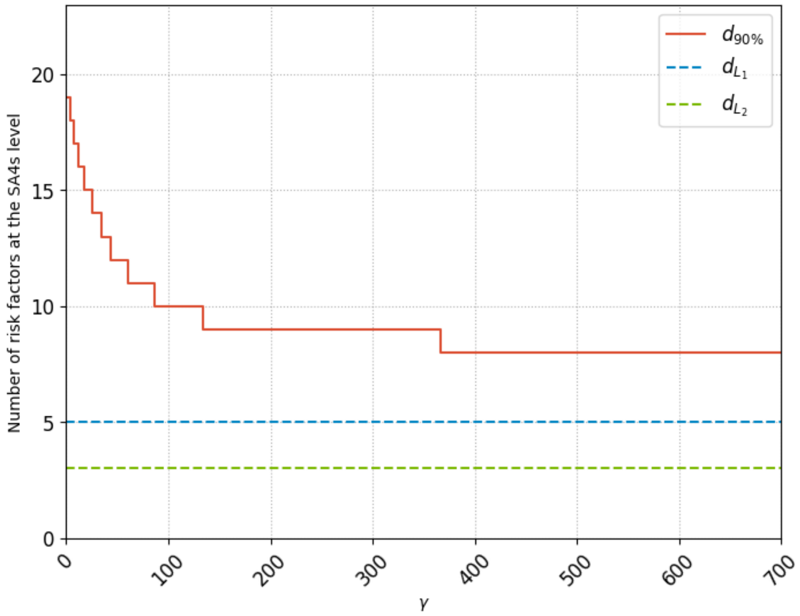
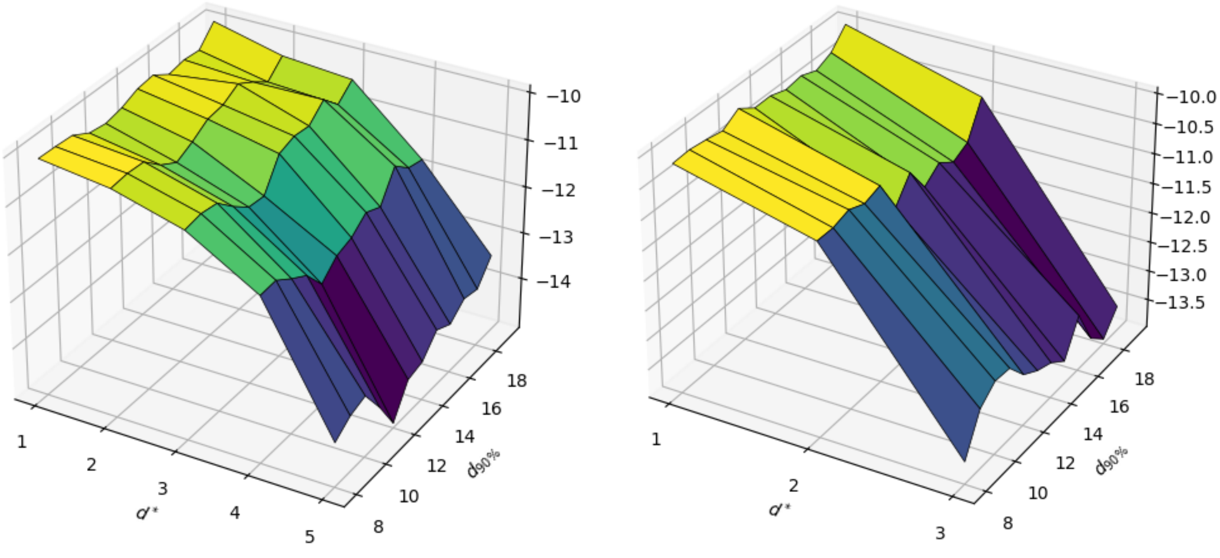


Fig. 3. The relationship between the number of risk factors and penalty term γ

The validation method is used to tune 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, as well as a method for constructing risk factors that is consistent 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 3 demonstrates that, as γ increases, $d_{90\%}$ gradually decreases. 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 (21). However, incorporating all principal components as risk factors is not always the optimal approach, as it can increase the complexity of the model. The average Akaike Information Criterion (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 limitations of the model, specifically, the inability to perform Cholesky decomposition when the number of risk factors becomes too large. This characteristic is utilised to narrow the search range for the optimal of risk factors. Figure 3 illustrates the changes in the maximal number of risk factors as γ varies. The red line shows the trajectory of $d_{90\%}$, while the blue and green lines denote the upper bounds of the number of risk factors required to explain all areas 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 region. Therefore, for each gamma value, the best risk factors will be selected based on Recursive Feature Elimination (RFE) from $d_{90\%}$, with d_{L_1} (d_{L_2}) factors selected for each region. The indices of the risk factors corresponding to the γ will be selected for the regions ranked highest among all regions. Iterating over all values of γ enables the calculation of the global minimum of the AIC and determines the optimal risk factors and the lag parameter of VAR models. It is worth noting that the best number of risk factors selected using RFE may still be less than d_{L_1} (d_{L_2}). This could happen if there are not enough informative risk factors or if the relationships between the risk factors are not strong enough to be captured by the model.



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

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

Fig. 4. 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^*.

The findings derived from Figures 4(a) and 4(b) suggest that the VAR(1) model with 5 risk factors corresponding to γ values in the interval $[44, 69)$ is the optimal choice. The plots also demonstrate that increasing the number of selectable risk factors leads to a better model performance based on the information criterion. However, it is important to note that the number of risk factors that can be included is limited by the characteristics of the VAR model and the SA4-level data, which exhibit low frequency and small data volume.

3.4 Residuals from the hierarchical model

The residual of the two-level hierarchical house price model for excess HPI growth rate of the i^{th} suburb is denoted as $e_{i,t}(\Omega_j)$, which the model cannot explain at granular scales due to the limitation of data collection.

Proposition 3.2. *The lack of information will lead to inaccurate estimates of the 2-level hierarchical house price models. The difference between the mean value of house prices and forecasts are*

$$\begin{aligned}
&\mathbb{E}_\omega\left[h_{i,t}\right] - \mathbb{E}_\omega\left[h^P_{i,t}\right] \\
&= \mathbb{E}_\omega\left[\mathbf{w}_i\right]^\top \mathbb{E}_\omega\left[\boldsymbol{f}_{i,t}\right] \sum_{\kappa>0} L^\kappa - \mathbb{E}_\omega[\mathbf{w}^P_i]^\top \mathbb{E}_\omega\left[\boldsymbol{f}^P_{i,t}\right] \sum_{\kappa>0} L^\kappa \\
&- \mathbb{C}\mathrm{ov}_\omega\left(\mathbb{E}_\omega\left[\mathbf{w}^P_i \mid I_{i,t-1}(\Omega, \Omega_j, \Omega_{ji})\right], \mathbb{E}_\omega\left[\boldsymbol{f}^P_{i,t} \sum_{\kappa>0} L^\kappa \mid I_{i,t-1}(\Omega, \Omega_j, \Omega_{ji})\right]\right),
\end{aligned} \tag{26}$$

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

$$e_{i,t}(\Omega_j) = \mathbb{E}_\omega\left[h_{i,t}\right] - \mathbb{E}_\omega[h^P_{i,t}] - \epsilon_t - \epsilon_{j,t} - \psi_t - \psi_{j,t} + \tilde{\epsilon}_{i,t}, \tag{27}$$

where $\tilde{\epsilon}_{i,t}$ is an unspecified random noise with mean 0 and finite variance, which describes the deviation of the expected house price growth index of the true model and the real house price growth index.

Proof. The lack of information can be interpreted as the lack of available data at each scale. According to Lemmas 3.1, 3.2, 3.3, and the introduction of lag terms, the difference between $\mathbb{E}_\omega\left[h_{i,t}\right]$ and $\mathbb{E}_\omega\left[h^P_{i,t}\right]$ is the error resulting from the lack of information. The final error of the hierarchical model is defined as

$$e_{i,t}(\Omega_j) = h_{i,t} - h^P_{i,t} - \psi_t - \psi_{j,t} = \mathbb{E}_\omega\left[h_{i,t}\right] + \tilde{\epsilon}_{i,t} - \left(\mathbb{E}_\omega[h^P_{i,t}] + \eta_t + \eta_{j,t}\right) - \psi_t - \psi_{j,t}. \tag{28}$$

Because VAR(1) models are selected both at the national and SA4 levels, residuals η are equivalent to the stochastic disturbances ϵ, which are noises with mean 0 and finite variance. □

Equation (27) in Proposition 3.2 indicates that the final errors of the hierarchical model in each area are time series, and the linear stochastic difference equation only consists of two distinct parts: a time-variant trend term and a combination of random noises. With the absence of a stationary component in the expression of the stochastic difference equation, the analysis of volatility becomes the focus of the next, after the series of residuals is detrended, which is regarded as a series of new observations in each area at the most granular scale.

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. According to the Augmented Dickey-Fuller (ADF) test, which tests whether univariate time series data is stationary or not, the proportion of series rejecting the hypothesis of being nonstationary is 0.9869 (454 out of 460) at the 1% level of significance. Therefore, $\mathbb{E}_\omega\left[h_{i,t}\right] - \mathbb{E}_\omega[h^P_{i,t}]$ is no longer time variant in this model, which can be interpreted by a constant term $\boldsymbol{\beta}^P_i(0)$ in the final model, only changing across different areas, illustrated as Equation (29).

$$\begin{aligned} &\text{HPI of Suburb i} \in \Omega_j : h^P_{i,t}(\Omega_j) = l_{0,t} + l_{j,t} + l_{ji,t}, \\ &D\left(\Omega_j, \Omega_{ji}\right) : l_{ji,t} = \boldsymbol{\beta}^P_i(0) + \boldsymbol{\beta}^P_i(\Omega)^\top \boldsymbol{f}^P_t(\Omega) \sum_{\kappa>0} L^\kappa + \boldsymbol{\beta}^P_i(\Omega_j)^\top \boldsymbol{f}^P_t(\Omega_j) \sum_{\kappa>0} L^\kappa. \end{aligned} \tag{29}$$

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

$$h^P_{i,t}(\Omega_j) = l_{0,t} + l_{j,t} + \boldsymbol{\beta}^P_i(0) + \boldsymbol{\beta}^P_i(\Omega)^\top \boldsymbol{f}^P_t(\Omega) + \boldsymbol{\beta}^P_i(\Omega_j)^\top \boldsymbol{f}^P_t(\Omega_j). \tag{30}$$

3.5 Volatility analysis in each suburb

This section provides a detailed analysis of the residuals of the expected house price growth index of the actual model $\tilde{\epsilon}_{i,t}$, defined in Proposition 3.2. As $\boldsymbol{\beta}^P_i(0)$ has been moved in the expression of the final model, Equation (27) can be rewritten as

$$\tilde{\epsilon}_{i,t} = e_{i,t}(\Omega_j) + \psi_t + \psi_{j,t} + \epsilon_t + \epsilon_{j,t}, \tag{31}$$

where the right-hand side of Equation (31), which is equal to $h_{i,t} - \mathbb{E}_\omega[h^P_{i,t}]$, measures the difference between the actual and predicted house price index growth rates. This equation also illustrates the forecasting process through simulations, where the distances l_0 and l_j are estimated without the random term, and the unbiased estimation for each suburb is obtained from the hierarchical model. The simulated value for each suburb comprises an unbiased estimation and a simulated residual, and the values of l_0 and l_j are updated to the appropriate average of the simulated house price growth rate in each suburb.

Despite the stability of the distributional properties of observed residuals, their distribution in different areas is diverse and may not follow a normal distribution. Under the assumption that the dependence structure and the marginal cumulative distribution functions of residuals are stable, multi-dimensional empirical copulas will be used to generate pseudo-random samples. Multi-dimensional empirical copulas can be used to simulate errors by modelling the dependence structure between the residuals in different suburbs. This involves fitting an empirical copula function to the observed residuals, which provides a way to estimate the joint distribution of the residuals. Once the empirical copula has been fitted, it can be used to generate pseudo-random samples that capture the dependence structure of the residuals. These simulated scenarios can be used to generate forecasts of future house price growth rates at different geographical levels, which can help understand the uncertainty associated with the predictions.

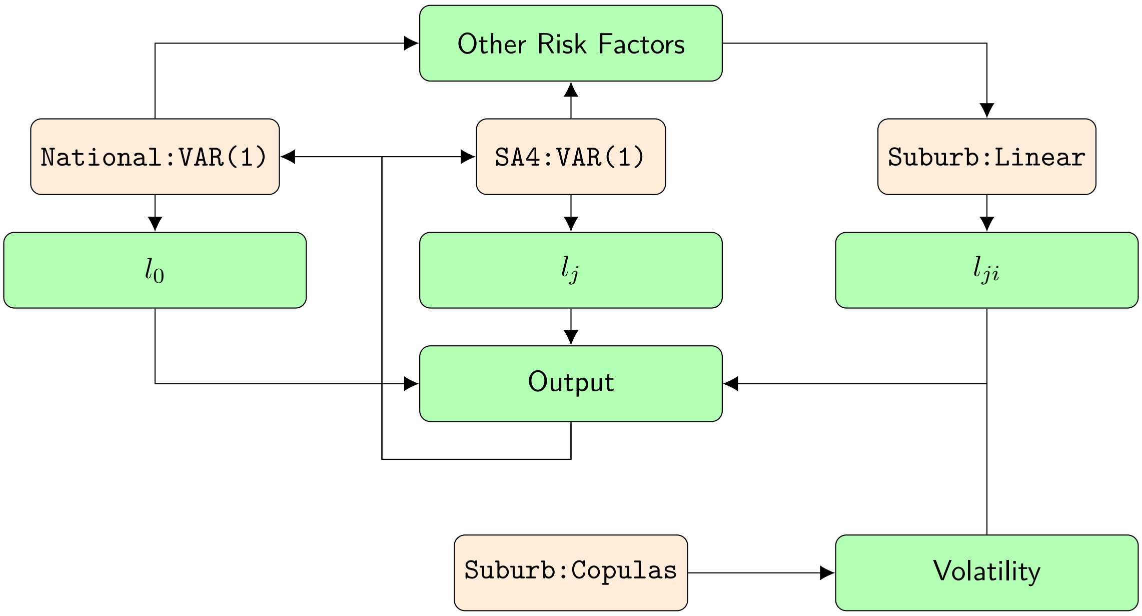


Fig. 5. The flow chart illustrates the simulation process employed for forecasts incorporating volatility analysis.

Figure 5 depicts the flow chart of the simulation process for forecasting with volatility 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 house price index growth rates for the country and SA4 level, in addition to an expected value derived from the linear model of the obtained risk factors. When volatility analysis is incorporated into the model, as shown in Figure 5, the unbiased estimate is augmented with stochastic errors generated based on multivariate copulas. The simulation outputs the final results of the forecast in one period, which include the errors in predicted average house price index growth rates at the country and SA4 levels. The values of l_0 and l_j are then obtained by averaging all the simulated results, and the corresponding error terms are computed.

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 $l_0$ 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 $\mathbb{E}[l_{0,t+1}]$ and $\mathbb{E}[l_{j+1}]$ based on VAR models.
          Estimate $\mathbb{E}[h^P_{i,t+1}(\Omega_j)]$ according to Equation (29) 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
```

The algorithm shown in Figure 5 is presented in more detail in Algorithm 1. The average of results from multiple simulations is considered the forecasted value for each time. The rationale behind integrating empirical copulas into the complete hierarchical model will be demonstrated in the next section.

4 Numerical results

The purpose of this section is to provide a comprehensive evaluation of the hierarchical model's performance. Firstly, the section simplifies the model and excludes the impact of external variables to allow for a more precise understanding of its fundamental components and mechanisms. Then, the predictive accuracy of the simplified model is assessed by comparing it to a similar predictive model, revealing valuable insights into the model's strengths and limitations. Furthermore, the impact of volatility analysis on the hierarchical model's outcomes is discussed, enabling a better understanding of the factors that may affect the model's results and facilitating further analysis of the risk factors related to house prices.

Moreover, the section compares the strengths and weaknesses of the hierarchical model and the

FAVAR model in interpreting and predicting house price index data, providing a comprehensive overview of their relative performance in different research contexts. Additionally, the advantage of using an RP-PCA model is demonstrated, followed by a sensitivity analysis of the RP-PCA method.

Finally, the section conducts an in-depth analysis of the risk factors obtained from both national and SA4 levels, shedding light on the characteristics of these variables and the sources and nature of the risks associated with the model. Moreover, we use a sensitivity analysis of the final model to provide an economic interpretation. This analysis establishes a solid foundation for fully comprehending the hierarchical model's performance and its suitability for various applications.

4.1 Comparing the hierarchical model and 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 $l_{0,t}$ and $l_{j,t}$ are estimated using ARIMA models. It is worth noting that the term $l_{ji,t}$ is set to be a constant over time, indicating the absence of any risk factors other than house prices. The residuals obtained from the hierarchical model are subjected to a volatility analysis to evaluate their suitability for generating reliable predictions. 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 errors. To justify the effectiveness of the simplified hierarchical model and its associated volatility analysis, it is compared to the MinT(Shrink) model proposed in Wickramasuriya et al. (2018). The MinT(Shrink) model is a predictive model which also utilises a hierarchical structure to forecast time series data. Similar to the hierarchical model, the MinT(Shrink) model first divides the time series data into groups and fits separate time series models to each cluster. The resulting models are then combined hierarchically to generate forecasts for the entire time series. Additionally, both models consider the influence of one-step errors in the base forecasts.

In order to compare the predictive accuracy of the hierarchical model and the MinT(Shrink) model, a rolling window approach is utilised. 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 - are evaluated over time using the root mean squared error (RMSE) as a measure. The MinT(Shrink) model is built upon the base model, while the hierarchical model (without volatility analysis) serves as the foundation for the hierarchical model (with volatility analysis). In the univariate case, the base model characterises the dependent variable's average, while the hierarchical model emphasises the distance of the average across various levels. At the lowest level, the hierarchical model uses a linear model to incorporate exogenously collected risk factors, which is a constant in the univariate case, whereas the base model still relies on ARIMA models to describe the objectives. The hierarchical model (with volatility analysis) differs from the MinT(Shrink) model in that it employs a numerical method that integrates the one-step error to create a numerical model. However, the use of a numerical approach in the hierarchical model is accompanied by 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 added to the hierarchical model cannot provide such theoretical guarantees with its numerical approach. The errors for 1- to 12-step-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 obtained through ARIMA models fitted using default settings in the automated algorithms proposed in Hyndman and Khandakar (2007) and implemented in the `statsmodels` package for Python. The outcomes of both the base and MinT(Shrink) models are validated through the utilisation of the `hts` package in R (Wang, 2021).

Tab. 5. Out-of-sample forecast performances of four modes at the postcode level

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

† Base line shows the average RMSE ($\times 10^{-2}$) of the base forecasts. A negative (positive) entry above this row shows the percentage decrease (increase) in the average RMSE of forecasts relative to the base forecasts.

As observed from Table 5, compared to the base forecasts, the MinT(Shrink) model shows an increase in the average RMSE for all steps ahead, with a maximum increase of 10.89% for the 1-step-ahead forecast. On the other hand, both the hierarchical model without empirical copulas and the hierarchical model with empirical copulas show a decrease in the average RMSE for all steps ahead. It is worth noting that the hierarchical model with empirical copulas consistently outperforms the MinT(Shrink) model and the hierarchical model without empirical copulas in terms of average RMSE for all steps ahead. The observed increase in the average RMSE of the MinT(Shrink) model for all steps ahead indicates that the model's performance deteriorates as the forecast horizon increases. This could be due to the model's inability to capture the complex dependencies among the residuals under the Gaussian assumption over longer time periods. On the other hand, the consistent decrease in the average RMSE of the hierarchical model (with and without empirical copulas) for all steps ahead suggests that these models are better suited for longer-term forecasting than the MinT(Shrink) model. This could be due to the hierarchical model's ability to capture the hierarchical structure of the data and incorporate information from higher levels when predicting values. Furthermore, the superior performance of the hierarchical model with empirical copulas compared to both the MinT(Shrink) model indicates that the inclusion of copulas in the model helps to capture the dependence structure of the data more accurately, resulting in improved forecasting accuracy. 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.

4.2 Results of the volatility analysis of the full hierarchical model

The numerical results presented in this section demonstrate the rationale behind integrating empirical copulas into the complete hierarchical model. Although the distributional properties of the residuals of the one-step-ahead forecasts of the full hierarchical model are stable, their distribution in different areas is diverse and may not follow a normal distribution. The Kolmogorov-Smirnov Test and Shapiro–Wilk test reject the null hypothesis that residuals follow a normal distribution at the 1% level of significance in every suburb. Figure 6 displays the distributions of residuals in six randomly chosen suburbs, arranged in ascending order of skewness from top to bottom. These distributions do not exhibit the characteristics of a normal distribution, consistent with the statistical tests. Furthermore, an increase in kurtosis is noticeable with an increase in skewness.

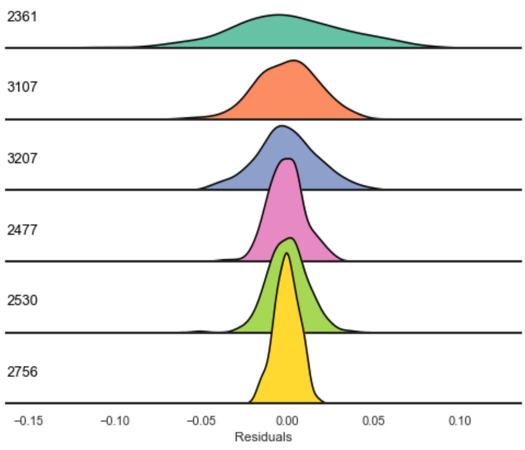


Fig. 6. Distributions of Residuals

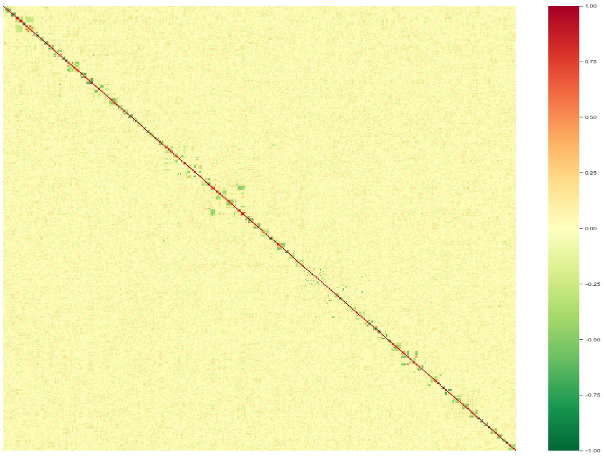


Fig. 7. Correlation Heatmap of Residuals

The correlation heatmap of residuals across all suburbs is displayed in Figure 7, with the grids arranged according to the postcodes of the suburbs. The similarity in postcodes is assumed to represent the geographical distance between the suburbs. A strong correlation tends to exist in suburbs with more similar postcodes. A noteworthy observation is that only the residuals in adjacent suburbs show positive correlation coefficients, as depicted by the distribution of tiny knots on the diagonal in Figure 7. In contrast, the correlation between residuals in less adjacent suburbs is negative, as shown by the squares of green grids distributed in pairs symmetrically around the diagonal. The correlation between more distant areas tends to be zero.

Figure 8 presents a comparison between the new simulation of l_0, taking into account the dependence structure of residuals, and historical values by displaying the probability dnesity function (PDF) of empirical distributions. This comparison indicates that the minor shift observed in the simulated results in Figure 2 has been resolved, thus substantiating the underlying assumptions concerning the volatility analysis.

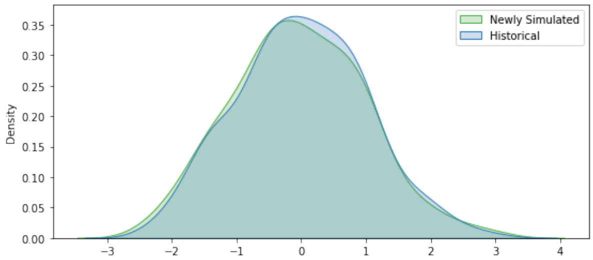


Fig. 8. New simulation results of house price index growth rate at the national level

Following this, simulations are conducted to estimate the expected shortfalls of HPI at the postcode level across various confidence levels. Figure 9 displays the relative changes between the expected shortfalls simulated with and without volatility at 5 years in the future (December 2025), based on the HPI in December 2020. The figure includes three graphs in category (a), depicting the results obtained using volatility analysis at three confidence levels, and three graphs in category (b), presenting the results obtained without volatility analysis. It is observed that category (b) tends to yield higher house price estimates in scenarios where house prices are low. Furthermore, the predicted shortfalls in category (a) exhibit greater fluctuations across different areas compared to those in category (b). One possible explanation for the observed pattern is that the capture of the stochastic term with the use of volatility analysis alleviates the bias in the distribution of the original simulated average house price index at the country level. In addition, country-level average house prices are heavily influenced by historical house prices, so minor errors in the simulation may cause significant biases.

In general, the reason for using multiple copulas is that the residuals obtained from the hierarchical model are interdependent among different postcode-level areas and the marginal distributions do not follow a normal distribution. By inserting multiple copulas, the RMSE of long-term predictions in different postcode-level areas can be reduced, as evidenced by Table 5, and the prediction performance for higher levels, such as the national level, can be improved. Additionally, when predicting scenarios in the long-term future, using volatility analysis results in more conservative forecasts.

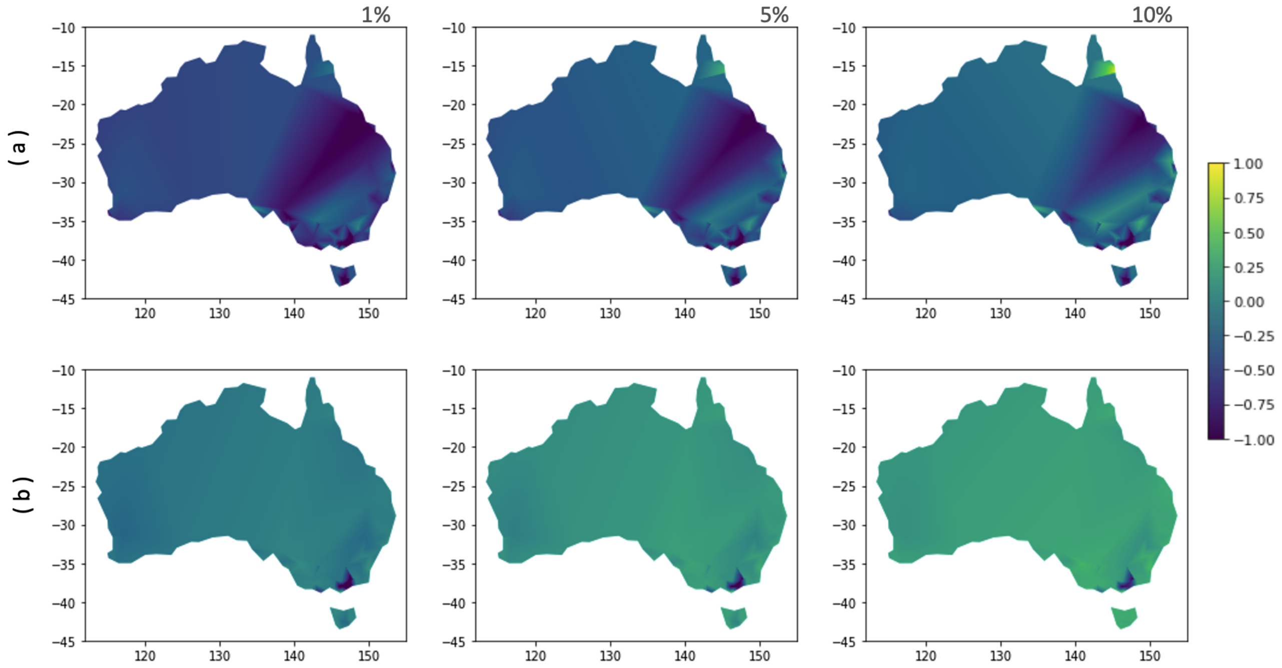


Fig. 9. Relative expected shortfall changes at 99%/95%/90% confidence levels

The aim of volatility analysis is to improve forecast accuracy rather than to provide economic interpretation. Therefore, the next two sections will exclude the analysis of volatility and instead concentrate on how the risk factors contribute to improving the goodness of fit and whether it is reasonable to incorporate these factors into the hierarchical model.

4.3 Comparing the hierarchical model and the FAVAR model

This subsection presents a detailed comparison between the hierarchical and FAVAR models, based on measures of goodness of fit, model complexity, and prediction accuracy. The FAVAR methodology is considered as a potential factor models to solve the limited information problem, as it combines standard structural VAR analysis with dynamic factor analysis for large data sets (Bernanke et al., 2005). Dynamic factor models propose that information from a considerable number of observations can be summarised by a relatively small set of estimated indexes or latent factors.

The available number of house price index growth rates in area j 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 house price index 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 $|\Omega_j| \times M$, and the $|\Omega_j| \times 1$ vector of error terms ξ_t are mean zero and assumed normal and 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 analyse1, 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 for the growth rates of the all-house price index 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 amount of cross-correlation. The joint dynamics of $\left(\boldsymbol{f}_t^L(\Omega), \boldsymbol{f}_t^P(\Omega)\right)$ is expressed as:

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

Three subtypes of Hierarchical models and the two FAVAR models are compared from the perspectives of model construction and explanatory power and the comparison is shown in Table 6. These three subtypes are simplified hierarchical, half hierarchical, and full hierarchical models, with the simplified Hierarchical model already introduced in the section "Comparing the Hierarchical Model and MinT(Shrink) Model". It is worth noting that, an AR model is used at this point to maintain consistency with the other two hierarchical models. The half-hierarchical model refers to a model where, at the SA4 level, risk factors are constructed according to the special case 1 of the RP-PCA model according to Equation (21).

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 contains similar components, consisting of 13 risk factors. However, in different areas, the risk factors in the FAVAR2 model are the same at each time point, whereas in the FAVAR1 and the hierarchical models, these risk factors can vary between areas at the same time point. The goodness of fitting for different models is represented by the prediction accuracy on the training data, which is the average training RMSE of different rolling windows with a length of 120. From Table 6, it can be observed that an increase in model complexity leads to an increase in the degree of data interpretation, which is consistent with common sense. By comparing the Half and Full Hierarchical models, it can be observed that the use of the RP-PCA model can significantly increase the model's interpretive ability.

Tab. 6. Comparing FAVAR models and Hierarchical models for goodness of fit.

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

[†] In the columns of National Risk Factors 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 housing price data itself, multiplied by the total number of regions in that particular hierarchy c.

Table 7 provides evidence that incorporating more risk factors into the hierarchical model does not significantly compromise the accuracy of future forecasts. 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. This suggests that the introduction of external data through RP-PCA can help reduce errors in long-term forecasting. Furthermore, 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. In contrast, the FAVAR2 model used for comparison exhibits superior predictive performance in the short term, but the errors gradually increase over time. One possible reason is that the FAVAR2 model is oversimplified, which leads to good forecasting performance but limited explanatory power. This viewpoint is also supported by Table 6. It is worth noting that the full hierarchical model outperforms the half-hierarchical model in both the shortest and longer forecasting periods. Therefore, Table 7 supports the use of the RP-PCA model in the hierarchical model and affirms its credibility in providing improved forecasts for house price data.

Tab. 7. Out-of-sample forecast performances of five-factor models at the postcode level.

	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

[†] 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 section compares two FAVAR models and three hierarchical models and shows that incorporating risk factors through the RP-PCA model does not significantly undermine the models' predictive ability. Moreover, including risk factors improves the models' economic interpretability. Given that the external SA4 level data used is annual, the results obtained are reasonable and encouraging.

4.4 Robustness of the adapted RP-PCA

Tab. 8. Loadings of risk factors at the SA4 level.

	Risk Factor 1	Risk Factor 2	Risk Factor 3	Risk Factor 4	Risk Factor 5
hsdebt	-0.047070	-0.054248	0.084555	-0.374089	0.080772
wsce	0.044069	0.046119	0.186418	0.052046	-0.191518
agemedian	-0.285835	-0.051730	0.061505	-0.169447	-0.280327
agestd	0.422337	-0.015706	0.240085	0.012874	0.104675
ageq3	0.128121	0.065043	-0.166823	-0.273128	0.422180
agecount	-0.152731	-0.009628	-0.014021	-0.080853	-0.095862
xphmrna	0.118330	-0.116926	0.496790	-0.054269	-0.148198
xphltpa	0.251650	-0.645265	-0.267337	-0.373234	-0.081543
lsemp	-0.222423	-0.499974	0.043077	0.374620	0.413543
iprbwm	0.135224	0.355165	0.074928	0.064313	0.054593
fiprbmr	0.104659	0.033791	-0.070040	-0.021481	0.048690
ancob_1101	-0.081477	0.102213	-0.157588	0.075352	0.072681
ancob_1201	0.149737	0.093088	0.133354	0.169603	-0.025108
ancob_7103	-0.210592	0.170861	0.108802	-0.092113	-0.003586
ancob_6101	-0.265973	-0.039421	-0.002397	-0.187405	-0.071859
ancob_5105	0.205265	-0.046733	-0.116374	0.142440	0.148147
edhigh1_4	-0.065014	0.212562	-0.620584	0.098875	-0.102674
edhigh1_5	-0.132713	-0.072435	-0.029292	-0.097503	-0.316257
edhigh1_8	-0.020149	-0.090199	-0.124664	0.056944	0.054345
edhigh1_9	-0.178908	-0.116395	0.157326	-0.073592	-0.116595
anatsi_2	-0.267808	-0.187428	0.073847	0.481918	0.018516
anatsi_3	-0.477178	0.099848	0.079636	-0.246737	0.345361
chkb12_1	0.005692	-0.105506	-0.178688	0.201853	-0.441687

† the descriptions of all these variables are listed in Table 2.

In a fitted RP-PCA model, the loadings of risk factors at the SA4 level provide insights into the contribution of original variables to each component, as shown in Table 8. To test the sensitivity of the model, a simulation approach is used. In each simulation, 5 SA4-level time series generated randomly from a standard normal distribution are assumed to represent the SA4-level average. For each SA4-level area and variable, a newly generated time series also subject to a standard normal distribution is added to the known average. The resulting time series is then inserted into the original dataset. Multiple simulations are conducted, and the resulting loadings are averaged to generate a new Table 9. By comparing the changes between Table 8 and Table 9, the stability of the RP-PCA method can be determined. The results indicate that the RP-PCA method is relatively stable and reliable.

Tab. 9. Robustness Test: Loadings of risk factors at the SA4 level.

	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
wsce	0.040468	0.044586	0.189539	0.053279	-0.193295
agemedian	-0.288531	-0.049618	0.060317	-0.164907	-0.277348
agestd	0.421709	-0.014769	0.241167	0.014759	0.105119
ageq3	0.127716	0.068639	-0.161414	-0.273830	0.424677
agecount	-0.153737	-0.010077	-0.013228	-0.079194	-0.095319
xphmrna	0.117442	-0.117297	0.498115	-0.057740	-0.149872
xphltpa	0.251271	-0.645910	-0.265797	-0.371495	-0.080600
lsemp	-0.222298	-0.498821	0.046789	0.372046	0.414242
iprbwm	0.133836	0.355005	0.077661	0.067055	0.055210
fiprbmr	0.106791	0.034247	-0.069302	-0.021411	0.048765
ancob_1101	-0.083466	0.100582	-0.158472	0.073831	0.072029
ancob_1201	0.148425	0.090570	0.133342	0.167515	-0.025991
ancob_7103	-0.208439	0.169382	0.108600	-0.093309	-0.005138
ancob_6101	-0.264256	-0.039849	0.001015	-0.186877	-0.073736
ancob_5105	0.205965	-0.044643	-0.115827	0.143242	0.148128
edhigh1_4	-0.067241	0.212654	-0.621667	0.099669	-0.104623
edhigh1_5	-0.130216	-0.073308	-0.032706	-0.099762	-0.315203
edhigh1_8	-0.019074	-0.092077	-0.127864	0.060256	0.055505
edhigh1_9	-0.179356	-0.117898	0.156384	-0.075394	-0.114189
anatsi_2	-0.269164	-0.184877	0.071526	0.481738	0.019514
anatsi_3	-0.479766	0.100232	0.078930	-0.247900	0.344730
chkb12_1	0.004765	-0.105922	-0.179660	0.203384	-0.444280
SA_1	0.000536	-0.000117	0.002385	-0.000285	0.001253
SA_2	0.001308	0.001256	0.002847	-0.002580	0.001410
SA_3	-0.004404	-0.000404	0.001973	0.003849	0.000927
SA_4	-0.001619	0.004423	0.000447	0.001142	0.003052
SA_5	0.003230	-0.002180	-0.000744	0.003263	0.000414

† Below the dashed line are the average loadings of the newly introduced variables which are randomly generated

4.5 Analysis of risk factors

This section will focus on exploring potential factors that could account for the variation in historical house prices across different areas, and highlight some of the key findings from the model that could shed light on these factors.

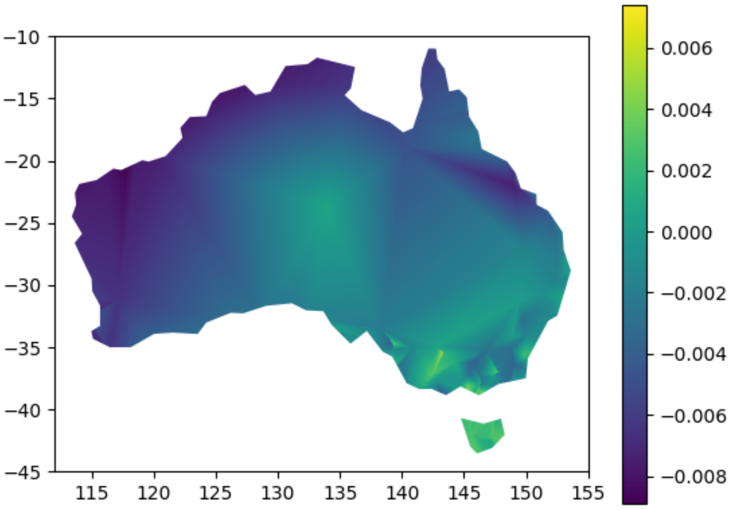


Fig. 10. The value of $\beta_i^P(0)$ in Equation (30) corresponds to each suburb.

Tab. 10. 10 suburbs with the highest values of $\beta_i^P(0)$ in Equation (30).

Postcode	Longitude	Latitude	Constant($\times 10^{-3}$)	Place Name
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

Tab. 11. 10 suburbs with the lowest $\beta_i^P(0)$ in Equation (30).

Postcode	Longitude	Latitude	Constant($\times 10^{-3}$)	Place Name
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

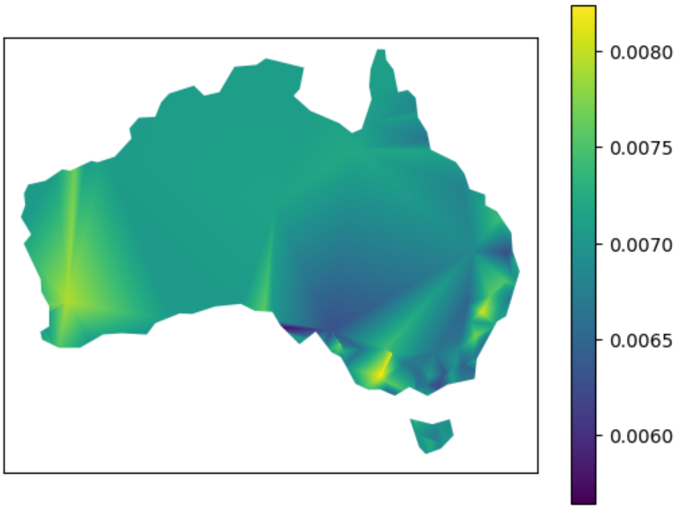


Fig. 11. Effect of previous l_0 on house prices in each suburb.

Tab. 12. 10 suburbs with the highest coefficients of l_0 in the previous period.

Postcode	Longitude	Latitude	Coefficient($\times 10^{-3}$)	Place Name
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

Tab. 13. 10 suburbs with the lowest coefficients of l_0 in the previous period.

Postcode	Longitude	Latitude	Coefficient($\times 10^{-3}$)	Place Name
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

The final hierarchical one-step-ahead forecasting model can be expressed as a linear combination of a constant term, the current national-level risk factors, and the current SA4-level risk factors, including l_0 and l_j. Figure 10 - 11, Table 10 - 12, and Table 13 show that house prices in different areas have varying fixed growth rates, which are represented by the constant terms in the final model. Suburbs closer to the southeast and big cities of Australia have higher fixed growth rates, while suburbs closer to the northwest and away from big cities have lower fixed growth rates, where house prices even experience negative fixed growth rates annually. However, some areas with lower fixed growth rates are heavily influenced by the overall national housing market. In other words, house prices in these areas may have experienced growth on average during these years but were largely driven by the overall housing market, which can be attributed to changes in the economic system, as well as the radiating effects of price increases in other areas [1].

At the SA4 level, there exists five exogenous risk factors and one l_j, which is the distance of fixed house price growth rates of the national and the SA4 levels. Despite the employment of a consistent construction method, the coefficients and values associated with these risk factors vary significantly across different time points. 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 as the contribution of risk factors to house price index changes in each area. Because all risk factors are generated by using principal components analysis, the partial effects are equivalent to the marginal effects. In year N, the proportion of the marginal effect of the annual change in a particular factor to the marginal effect of the annual change in the log of the national house price is calculated in order to determine the impact of that factor on changes in house prices

$$\frac{\text{coefficient}_f\left(\Omega_{ji}\right) \times \sum\limits_{t \in \text{Year } N} f_t^P\left(\Omega_j\right)}{\text{coefficient}_{l_0}\left(\Omega_{ji}\right) \times \sum\limits_{t \in \text{Year } N} l_{0,t}} \times 100\%, \tag{36}$$

which is considered the effect of the factor on housing price changes for in year $N+1$. 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_0 tends to remain at a relatively stable level, with the coefficient of l_0 all positive as illustrated in Table 13, making it a more suitable denominator for comparison. This constructed measure can also be considered a distortion of the average housing price caused by a particular factor at the SA4 level in a suburb. As shown in Figure 12, it is evident that the contributions of one risk factor vary significantly across different areas. However, due to the lack of clear economic interpretations of these risk factors, further explanations cannot be provided at present. Additionally, attempting to analyse the original variables of these risk factors is not preferred, considering the reason for using principal component analysis is to eliminate the possible correlations between different variables and cut down dimensionality.

[1] The coefficients or loadings of risk factors at the national level, along with the corresponding geographical distribution, are presented in the Appendix. A more in-depth analysis of these risk factors will be provided in conjunction with the SA4-level risk factors later

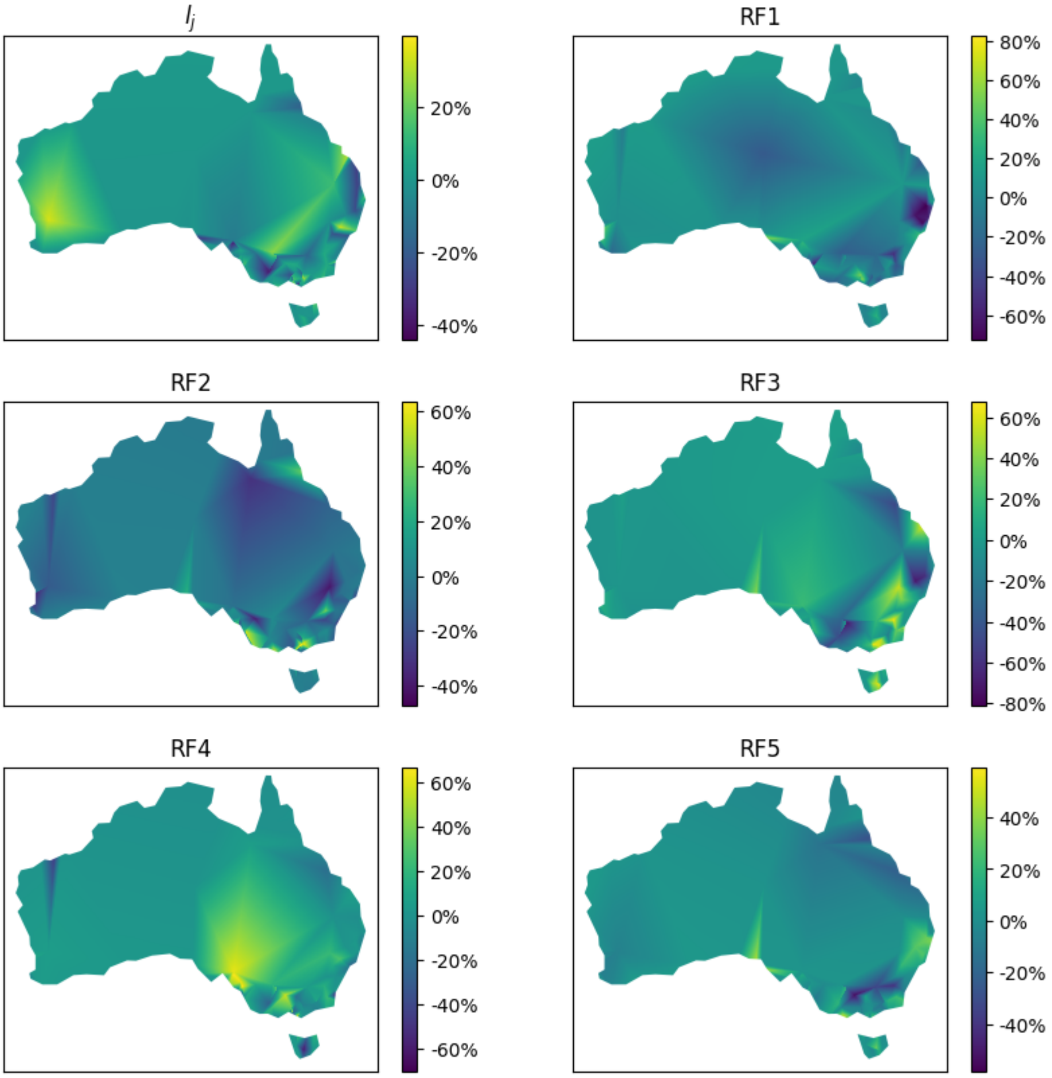


Fig. 12. The average historical contribution of risk factors at the SA4 level to the overall house price changes in each suburb. The average historical contribution of risk factors is defined in Equation (36).

Because these risk factors (excluding l_0 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. This expectation is supported by the historical dataset and is demonstrated in Figure 13, which displays the trend of average partial effects of the factors in the model over years in all suburbs.

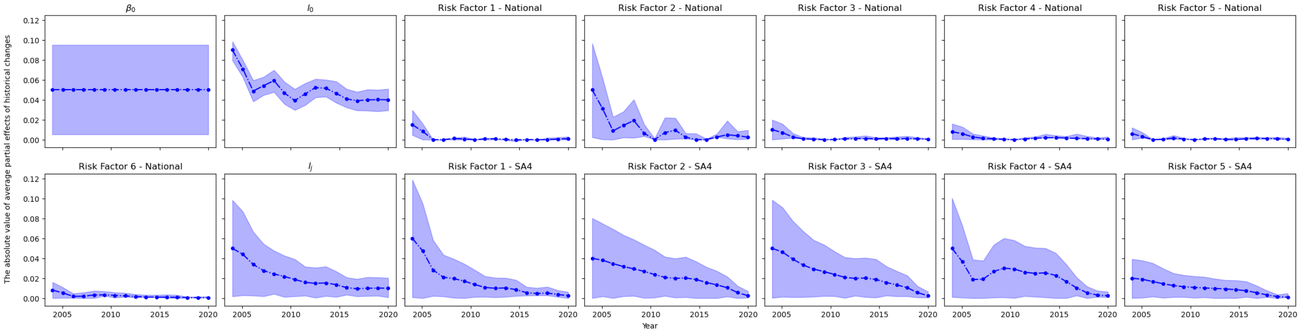


Fig. 13. The absolute values of average partial effects of historical changes over time in all suburbs. This figure illustrates the decreasing trend of absolute values in the average historical changes of risk factors over time. The line represents the mean value across all suburbs, while the interval encompasses 99% of the suburbs.

The purpose of including the constant l_0 and national risk factors in the plot is to aid readers in comprehending and comparing risk factors at different levels. The plot demonstrates that the average partial effects of historical changes in l_0 and l_j become gradually stable over time at a relatively constant level. In addition, the average partial effects of historical changes in national risk factors and SA4-level risk factors both tend towards zero. 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 the constant, l_0 and l_j terms will still be the most significant factors, while other risk factors will only have a cumulative impact in the short term. This aligns with the original intention of building a hierarchical model, which aims to provide reasonable explanations in one-step prediction and not compromise long-term forecasting results. The inclusion of the constant term and the hierarchical structure of the model enable it to capture the overall trends and patterns of the data, while the inclusion of risk factors allows for short-term fluctuations and changes to be taken into account.

4.6 Sensitivity analysis of the hierarchical model with volatility analysis

In this section, we will adjust a single preprocessed variable from Table 2, with the remaining variables fixed. The change is assumed to be an increase of one standard deviation from the initial value. We

Tab. 14. The average and standard deviation of predicted house price changes across different 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 achieved: Year 12 or equivalent	0.41%	4.42%
Highest education level achieved: Cert III/IV	1.62%	4.10%
Highest education level achieved: Bachelor degree	2.12%	3.29%
Highest education level achieved: Postgraduate degree	2.36%	3.98%

[†] The dashed lines separate the table into three parts: national-level variables, economic variables on the SA4 level, and geographic variables on the SA4 level.

observe how this one-unit change will influence the predicted house price index (average) in 5 years across different areas on the postcode level. Table 14 illustrates the average and standard deviation of predicted house price changes across different areas.

We start with the discussion on the national-level variables. The increase in interest rate growth elevates borrowing costs and decreases house demand. This effect will be more pronounced in suburbs in metropolitan areas such as Sydney and Melbourne. Exchange rate growth plays a significant role, too. A higher appreciation in the exchange rate results in a decline in foreign investment in the housing sector. Prime locations such as Sydney CBD, Pyrmont, Chatswood, Melbourne CBD, Box Hill, and suburbs in Golden Coast will be particularly sensitive. A surge in the Consumer Price Index growth, when not complemented by corresponding wage hikes, will generally suppress house demand. On the positive side, a unit increase in Gross Domestic Product growth will uplift house prices, reflecting the nation's economic robustness. This effect is directly observable in GDP-centric areas in New South Wales and Victoria such as North Sydney, Parramatta, Docklands and Clayton. An increment in 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 higher bullish swing in the ASX will spur the real estate market. Wealthier suburbs (Point Piper, Bellevue Hill, Toorak, South Yarra, East Melbourne, Ascot, and Hamilton) will reflect this trend more obviously.

Next, we analyse the economic variables on the SA4 level. A rise in home debt growth signifies two concurrent phenomena: increased buying activity, pushing house prices up (among suburbs in Sydney and Melbourne), or burgeoning financial stress which will later depress house prices (regional towns or natural landscapes). Therefore, the geographical variation of the influence of home debt changes is pronounced. An increment in wage growth typically bolsters purchasing power, potentially raising house prices. Wage-heavy suburbs in major cities will be especially responsive. A surge in home improvement activities generally insinuates housing value appreciation with minor geographical variations. Fees paid to health practitioners are a proxy for health-conscious locales (Bellevue Hill, Vaucluse, Toorak, and Ascot), where a larger rise is correlated with pricier 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 and could not pay the mortgage or rent on time) foreshadows future price corrections, with the most vulnerable regions such as parts of Western Sydney and outer suburbs of Melbourne potentially showing sharper downturns. A shift in employment dictates disposable income levels and borrowing capacities, generally pushing up house prices.

Next, we will examine the demographic-related variables on the SA4 levels. Population growth typically exerts upward pressure on house prices, and areas with high population density (Surry Hills, Chippendale, Southbank, Fitzroy, Kangaroo Point, East Perth, and Bowden) tend to face more serious housing supply constraints, magnifying house price hikes. An interesting observation is that 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, Cert III/IV, Bachelor's, and Postgraduate), a burgeoning house market due to upper shifts in income brackets and housing preferences will be observed. Other variables such as the country of birth and indigenous status may suggest housing needs and financial capabilities. However, they have smaller influences compared to 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.

5 Conclusion

In the existing literature on describing and forecasting local house prices using the VAR model and its extensions, the VARX model introduced in Shao et al. (2015) is the only previous approach that attempts to use information dynamically from the granular levels to describe the geographical variation of house price growth rate. However, only house prices are treated as risk factors, which makes sense as other cumulative partial effects of risk factors at the granular levels have been testified to approach zero in the long run. However, this report also shows that some risk factors at granular levels highly influence the forecasts in the short term.

In order to incorporate these factors, this paper, based on the conditional CAPM model proposed in Jagannathan and Wang (1996), derives a hierarchical structure that provides a solid foundation for introducing external variables from different regions. To ensure the model's simplicity, the spatial application of the RC-PCA approach proposed in Lettau and Pelger (2020) is introduced, which recombines external variables into new lower-level risk factors and integrates them into the model.

The utilisation of information from a model's residual covariance structure has been demonstrated to enhance prediction accuracy, as evidenced by 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. A simulation-based approach that incorporates copulas of these residuals into the hierarchical model has been proposed and has been shown to improve the prediction accuracy of 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 predictive ability of the model. This preservation of predictive ability is jointly contributed by RP-PCA and residual analysis. By introducing these risk factors, the model's short-term prediction and explanatory power have been enhanced to some extent. This allows decision-makers to quickly adjust their predictions for the next stage of house prices in a region when observing changes in certain factors across different areas. 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 housing price risks.

In this model, aside from risk factors, other coefficients are assumed to be constant, but this may not be the case in reality. Moreover, the residual analysis only considers one-step residuals, which not only refer to the temporal aspect but also the spatial aspect, as some theoretical derivations in this paper transform conclusions from time to space. As a result, it is worth further exploring whether the final residuals are related to the residuals generated between different levels in the hierarchical model. This relationship may affect the model's stability and predictive ability, so it is necessary to fully consider this relationship in practical applications. In future research, researchers can explore introducing dynamic coefficients to more accurately capture changes in house price risk, thereby further enhancing the model's predictive ability. At the same time, they can attempt to study the impact of multi-step residuals to better understand and address this issue.

Disclosure statement

The authors have examined potential sources of bias and hereby declare that there are no conflicts of interest pertaining to this research or its findings.

Acknowledgements

The authors acknowledge financial support from the Australian Research Council Centre of Excellence in Population Ageing Research (CEPAR) and UNSW Business School. The opinions, findings and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the aforementioned institutions or any of their affiliates.

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). The findings and views reported in this paper, however, are those of the author(s) and should not be attributed to the Australian Government, DSS, or any of DSS' contractors or partners. DOI: doi:10.26193/PI5LP.

The authors would like to acknowledge and express gratitude to CoreLogic for providing house price index data. 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.

References

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

Australian Bureau of Statistics (2022). Housing Occupancy and Costs. https://www.abs.gov.au/statistics/people/housing/housing-occupancy-and-costs/latest-release.

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.

Chomik, R. and Yan, S. (2019). Housing in an ageing Australia. https://cepar.edu.au/resources-videos/research-briefs/housing-ageing-australia-nest-and-nest-egg.

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

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

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.

Green, R. K. and Malpezzi, S. (2001). *A primer on US housing markets and housing policy*. Urban Institute Press.

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.

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. N. and Shim, I. (2012). Monetary Policy Surprises, Credit Costs, and Economic Activity. https://www.rba.gov.au/publications/confs/2012/kuttner-shim.html.

LeSage, J. P. (2008). An introduction to spatial econometrics. *Revue d'économie industrielle*, (123):19–44.

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.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

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

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

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.

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.

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

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

A Risk factors at the national level

Tab. 15. Loadings of risk factors at the national level.

	Risk Factor 1	Risk Factor 2	Risk Factor 3	Risk Factor 4	Risk Factor 5	Risk Factor 6
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

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 actually not needed. We utilise a framework comprising eight stationary time series labelled from `A` through `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%.

Table 16 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.

Tab. 16. Comparison of the MSE of predicted house prices from the VAR model incorporated with PCA on 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

† 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 17 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 explanatory power 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.

Tab. 17. Comparison of the MSE of predicted house prices from the VAR model incorporated with PCA on 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

[†] 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.

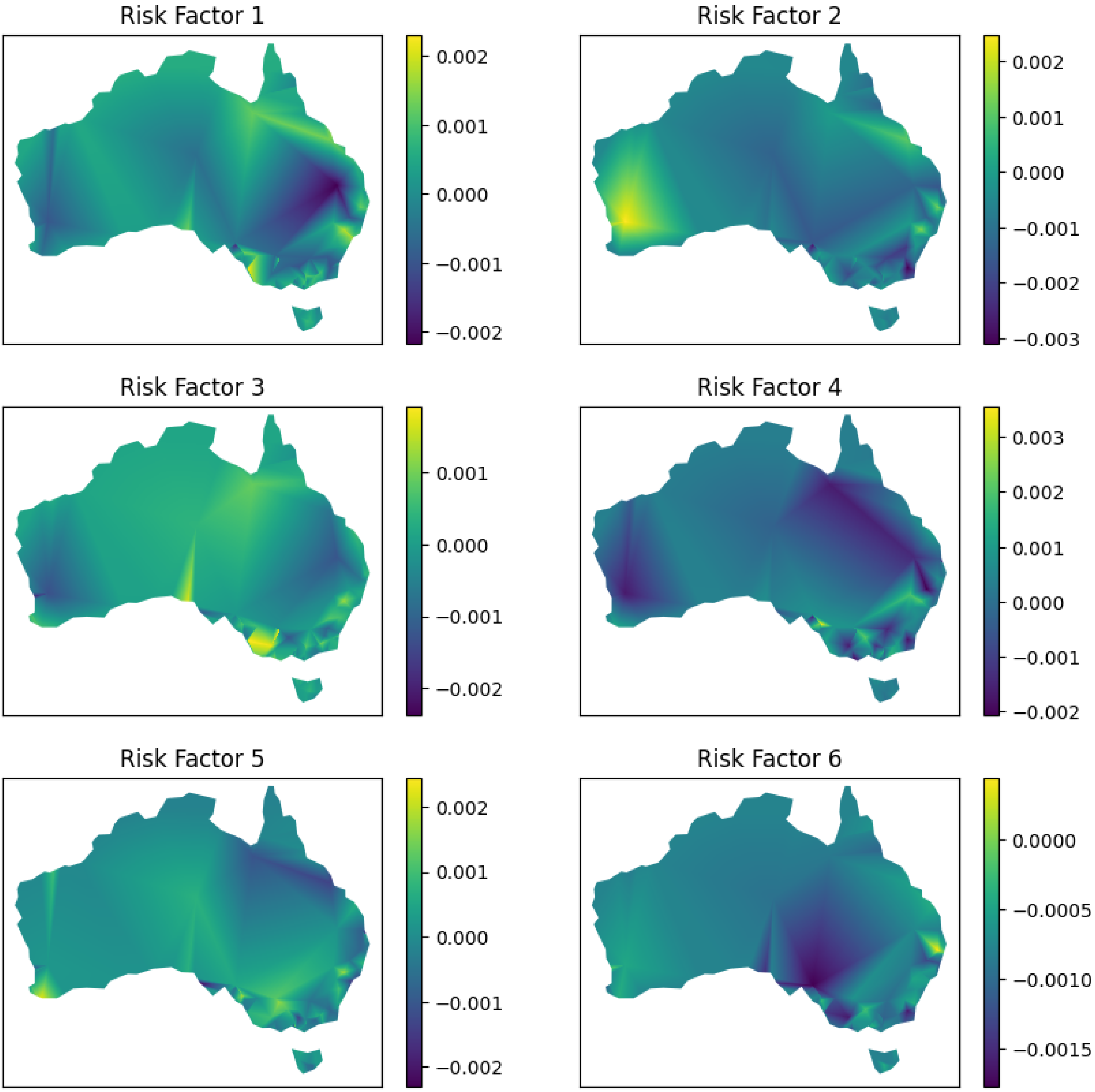


Fig. 14. Coefficients of national risk factors in each suburbs