Dynamics of housing markets

Tay, Darrell Jiajie

2018

Tay, D. J. (2018). Dynamics of housing markets. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/73251>

<https://doi.org/10.32657/10356/73251>

NANYANG
TECHNOLOGICAL
UNIVERSITY

DYNAMICS OF HOUSING MARKETS

TAY JIAJIE, DARRELL

SCHOOL OF PHYSICAL AND MATHEMATICAL SCIENCE

2017

DYNAMICS OF HOUSING MARKETS

by

TAY JIAJIE, DARRELL

Dissertation Supervisor: Associate Professor Cheong Siew Ann

Dissertation Co-Supervisor: Adjunct Professor Tee Shang You

A dissertation submitted to the
School of Physical and Mathematical Science
in partial fulfilment of the
requirements for the degree of
Doctor of Philosophy (Ph.D.)

School of Physical and Mathematical Science
Nanyang Technological University

August 2017

Abstract

It is known that contagion from housing markets crashes can cause large financial catastrophes such as the United States Subprime Mortgage Crisis leading to the Global Financial Crisis. Furthermore, housing markets are illiquid, making their recovery from crashes slower. While reviewing previous works on housing bubbles, we find that housing bubbles are often poorly defined. Also, these studies tend to each focus on a single factor. However, since housing markets consist of large number of heterogeneous agents with non-linear interactions, many have recognized the importance of studying them as complex systems. We find a Gibbs-Pareto equilibrium home price distribution for the markets we have looked at. We then identify bubbles as statistically significant deviations from these equilibrium distributions, providing a clear and unambiguous definition of bubbles. Next, we detected early warning signals for the bubble in the Singapore condominiums that started in 2007 in districts 9/10 before spreading out to the rest of the island. Similar early warning signals were also detected for the Hong Kong housing bubble that started in 2004. Lastly, we present correction formulas to improve parameter estimation and statistical significance when dealing with noisy and finite (housing) data.

Acknowledgements

Throughout this four-year PhD journey, I am gratefully to many people, without whom, this thesis could not have been possible. The first I would like to thank is my main supervisor, Assoc. Prof Cheong Siew Ann, who took me under his wing since my undergraduate days and taught me many valuable lessons, both in Life and in Science. To you, I owe a big thank you. My co-supervisor, Adj. Prof Tee Shang You, I am grateful for the opportunities you have given me and the projects we embarked on together. To Assoc. Prof Chew Lock Yew and Dr Christopher Monterola of my thesis advisory committee, thank you for your critiques, that made the works presented here more valuable and insightful. My collaborators, Prof Li Sai Ping, Prof Chou Chung-I, Prof Chihiro Shimizu and Prof Michael Wong, I am thankfully for the opportunity to work together.

My friends and colleagues at the Statistical Multi-Scale Simulations lab, Boon Kin, Woon Peng, Misha, James, Wen Yuan and Wan Chap, I am grateful for the fruitful discussion and the fun we shared along this four-year roller coaster ride. Especially to Teck Liang, thank you for the years of friendship since junior college, your encouraging words, gave me the optimism to deliver to my greatest potential. Cheers to my friends in Para Limes, my PhD would not have been half the fun without the food, laughter and mayhem.

Most importantly, I am extremely fortunate to have my family, and for the trust and believe that had placed in me. Mummy and Daddy, you have raised me well. My brother Gerald, thank you for always being there to hear my rants and complains. To my best friend and wife, Xue Ling, I am blessed to have you as my constant pillar of strength. Thank you for being that constant source of support and joy and happiness.

Lastly, I am grateful to Nanyang Technological University for my Research Scholarship, and for the sponsorship of this fruitful PhD journey.

Contents

Abstract	ii
Acknowledgements	iii
List of Tables	viii
List of Figures	ix
1 Introduction	1
1.1 Motivation	1
1.2 Literature Review	2
1.2.1 Econophysics - An Interdisciplinary Approach	4
1.2.2 Housing Markets as Complex Systems	6
1.3 Aim and Objectives	8
1.4 Organization	9
2 Bubbles are Departures from Equilibrium Housing Price Distributions: Evidence from Singapore and Taiwan	11
2.1 Introduction	11
2.2 Data and Methods	15
2.3 Results	18
2.3.1 Equilibrium Home Prices	18
2.3.2 Departures from Equilibrium as Housing Bubbles	22
2.3.3 Spatial-Temporal Dynamics of the Singapore Housing Market	23
2.3.4 Comparison between Singapore and Taiwan Housing Markets	24
2.4 Discussion	25

2.4.1	Income, Wealth, and Home Price	25
2.4.2	Bubble and Bubble-Like Features in Home Price CDF	28
2.4.3	Cause of the Singapore Housing Bubble	29
2.5	Conclusion	30
3	Early Warning Indicators in Housing Markets	31
3.1	Introduction	31
3.2	Data and Methods	34
3.2.1	Detrended Fluctuation Analysis	35
3.2.2	Non-detrended Methods	40
3.3	Results and Discussion	41
3.4	Conclusions	45
4	Finite Sample Corrections for Parameter Estimation and Signifi-	
	cance Testing	48
4.1	Introduction	48
4.2	Data	50
4.2.1	Taiwan Housing Data	50
4.2.2	Taiwan Income Data	50
4.2.3	SGX Normalized Returns	51
4.3	Reexamining Significance Testing for Empirical Distributions	52
4.3.1	CSN(i) MLE Estimation of α	53
4.3.2	CSN(ii) KS Distance	53
4.3.3	CSN(iii) Determining x_{min}	54
4.3.4	CSN(iv) Significance Testing	54
4.4	Finite-Sample Adjustments	55
4.4.1	Parameter Adjustment for Finite Largest Element	55
4.4.2	Test Statistic Adjustment for FLE	58
4.4.3	Adjustment for Finite Number of Elements	59
4.5	The Effects of Random Noise	59
4.5.1	Relation between Distribution Noise and Sample Size	61
4.5.2	Relationship between Distribution Noise and KS Distance	62

4.6	Application to Significance Testing	63
4.6.1	Significance Level for a Given Distribution	63
4.6.2	Fitting to Empirical Data	64
4.7	Conclusion	66
5	Conclusions	67
5.1	Summary	67
5.2	Future Works	68
5.2.1	Housing Locale Theory	68
5.2.2	Housing Portfolio Theory	69
5.2.3	Agent Based Simulations	71
	Bibliography	72
A	Housing in Singapore and Taiwan	88
A.1	Housing Types in Singapore and Taiwan	88
A.2	Year-by-Year plots of the Singapore Housing Market	89
A.3	Spatial and Temporal Heat Map of the Singapore Housing Market . .	93
B	Consumer Price Indexes	113
C	Properties of $U(0, 1)$ and its Ordered Statistics	117
C.1	The Probability Integral transform	117
C.2	Ranked Statistics of $U(0,1)$	119
D	Finite Sample Effects (FSE) Corrections	120
D.1	Mathematical Derivations Related to Power-Law and Exponential Distributions	120
D.2	Adjustment for Finite Largest Element	121
D.3	Theoretical Approximation of the Relative Error	122
D.4	Relationship Between KS Distance, Distribution Noise, and Sample Size	124
D.4.1	KS Distance and Sample Size	124

D.4.1.1	Analytical Solution Relating Distribution Noise to Sample Size	125
D.4.2	Distribution Noise and Sample Size	126
D.4.3	Correlation Between KS Distance and Distribution Noise . . .	128
D.5	Fits to Real Data	130

List of Tables

2.1	The p -test values for housing and income distributions in Singapore and Taiwan	19
B.1	The components of the CPI in Singapore. The base year for this index is 2014.	114
B.2	The components of the CPI in Taiwan. The base year for this index 2011.	115
B.3	The components of the CPI in HK. The base year for this index 2015.	116
D.1	Notations used in Chapter 4: Mathematical derivations related to Power Law (PL) and Exponential Distribution (EXP)	120

List of Figures

1.1	A schematic illustrating the approach to a critical transition.	7
2.1	A comparison between housing types in Singapore and Taiwan	15
2.2	The per-square-foot (PSF) home price CDFs in Singapore	20
2.3	The district/sector map with the upper quartile price of the different property types in Singapore.	21
2.4	The PSF CDFs of the different property types in the GTA	24
2.5	The wealth distribution of the top 50 richest person in Singapore and Taiwan	27
2.6	The CDF of the Taiwan wealth distribution	28
3.1	The procedure to perform empirical mode decomposition	36
3.2	The IMFs of the original time series	38
3.3	The Lag-1 Pearson ρ autocorrelation of the Singapore condominium housing segment for Mean Price and Volume	42
3.4	The half-max cumulative power, $\Omega_{1/2}$, of the Singapore condominium housing segment for Mean Price and Volume	43
3.5	Spectrogram of the Singapore condominium housing segment for Mean Price and Volume	44
3.6	The half-max cumulative power, $\Omega_{1/2}$, of the HK residential housing segment for Mean Price and Volume	45
3.7	The Lag-1 Pearson ρ autocorrelation of the HK residential housing segment for Mean Price and Volume	46
3.8	The mean residential transaction price of homes in HK for the period 1992 to 2010	47

4.1	p -testing and fits of 2012 to 2014 Taiwan home price, Taiwan lower-tail income and 2009 to 2016 Straits Time Index normalized return .	55
4.2	Relative estimation errors of $\hat{\beta}_{unadj}$ and $\hat{\beta}_{adj}$	57
4.3	The median KS distances for KS_{unadj} and KS_{adj}	58
4.4	Log-log plot of d_{KS} against N for different deciles	60
4.5	Illustration of the distribution noise we would measure if we sample 10 elements from $U(0, 1)$, rescaled such that the largest element becomes 1	61
4.6	Relationship between distribution noise d_{DN} and sample size N at different deciles	62
4.7	p -testing results for Taiwan per square foot home price, Taiwan lower-tail income, Straits Times Index normalized return before and after finite-sample adjustments	65
A.1	HDB CDF Year By Year Plots from 2000 to 2012	90
A.2	Condominium CDF Year By Year Plots from 1995 to 2014	91
A.3	Landed properties CDF Year By Year Plots from 1995 to 2014	92
A.4	HDB Heat Map from 2000 to 2002	94
A.5	HDB Heat Map from 2003 to 2005	95
A.6	HDB Heat Map from 2006 to 2008	96
A.7	HDB Heat Map from 2009 to 2010	97
A.8	HDB Heat Map from 2011 to Feb 2012	98
A.9	Condominium apartment Heat Map from 1995 to 1997	99
A.10	Condominium apartment Heat Map from 1998 to 2000	100
A.11	Condominium apartment Heat Map from 2001 to 2003	101
A.12	Condominium apartment Heat Map from 2004 to 2006	102
A.13	Condominium apartment Heat Map from 2007 to 2009	103
A.14	Condominium apartment Heat Map from 2010 to 2012	104
A.15	Condominium apartment Heat Map from 2013 to 2014	105
A.16	Landed properties Heat Map from 1995 to 1997	106
A.17	Landed properties Heat Map from 1998 to 2000	107
A.18	Landed properties Heat Map from 2001 to 2003	108

A.19 Landed properties Heat Map from 2004 to 2006	109
A.20 Landed properties Heat Map from 2007 to 2009	110
A.21 Landed properties Heat Map from 2010 to 2012	111
A.22 Landed properties Heat Map from 2013 to 2014	112
 B.1 The CPI of Singapore and its components	 115
B.2 The CPI of HK and its components	116
 D.1 Log-log plot of d_{KS} against N for different deciles with fits for A_{KS} and α_{KS} at different percentile	 125
D.2 Relationship between distribution noise d_{DN} and sample size N at deciles	126
D.3 Relationship between absolute deviation of distribution noise from expected value	127
D.4 Fits for A_{DN} and α_{DN} at different percentile.	127
D.5 Correlation between d_{KS} and d_{DN} at different sample sizes	129
D.6 Year by year Taiwan income data from 1981 to 1990	131
D.7 Year by year Taiwan income data from 1991 to 2001, fitted to power law distributions	132
D.8 Year by year Taiwan income data from 2002 to 2011, fitted to power law distributions	133
D.9 Year by year Taiwan income data from 1981 to 1990, fitted to expo- nential distributions	134
D.10 Year by year Taiwan income data from 1991 to 2001, fitted to expo- nential distributions	135
D.11 Year by year Taiwan income data from 2002 to 2011, fitted to expo- nential distributions	136
D.12 Year by year SGX normalized returns from 2009 to 2016, fitted to power law distributions	137

Chapter 1

Introduction

1.1 Motivation

Extreme events such as sudden spikes and dips are common phenomena observed in any commodities or asset classes. While these events can be caused by exogenous factors such as geopolitical and legislative reasons, endogenous factors such as positive feedback and herding [1, 2, 3] can also drive bubbles and crashes in markets. However, bubbles and crashes happen at different time scales across different markets. Liquid markets such as stocks and foreign exchange markets operate on much faster time scales as compared to housing markets which are less liquid [4] and have lower transaction volume. Typically, a liquid market takes a shorter time to recover from crashes since the time scales in these markets are faster compared to non-liquid markets. During the peak of the Global Financial Crisis, the Dow Jones Industrial Average (DJIA), which is commonly used as a benchmark for the US stock market, fell from a high of 14,093.08 in Oct 2007 to a low of 6626.94 points in Mar 2009, before recovering to the pre-crash peak in Mar 2013, a process spanning about 5.5 years. Comparatively, the non-seasonally adjusted House Price Index (HPI) of the United States fell from a previous high of 378.76 points during 1st Quarter 2007 to a low of 309.37 points in 1st quarter 2011 before recovering to the pre-crash peak in 3rd quarter 2016, taking almost 10 years to recover.

Slow recovery aside, housing market crashes also affect a large cross section of society. For most people, a house is likely to be the most expensive purchase in

their lifetime. This is not surprising because housing markets play a vital role in society and directly impact the livelihood of people [5]. Since most people usually take up a mortgage to support their home purchases, they need to set aside a fix portion of their incomes to repaying their loan obligations. As such, this portion of their income which could have otherwise been disposable income is reduced. This disposable income which drive the consumer economy, can be further reduced if interest rate on the mortgage increase (as in the case of the US subprime loans, when low introductory rates expire and subprime rates come into force). In serious cases, this might come to a point where repayment becomes impossible. In a rising housing market, homeowners who find themselves in such predicament can sell their homes, pay back the bank and hopefully pocket any balance profit after fees. However, when housing markets go south, selling their homes is no longer an option, because the proceeds of the sale is not enough to settle the mortgage with their banks. Such dynamics further slow down the recovery of housing markets. In anticipation of this, homeowner caught in such a position might also be forced into panic selling, putting downward pressure on prices.

In Japan, the collapse of the asset bubble in the early 1990s, led to a period of economic stagnation known as the ‘Lost Decade’ [6]. Similarly, the US subprime mortgage bubble which began in 2004 and popped in 2007, was largely considered a primer to the Global Financial Crisis [7]. Given the possible catastrophic effects arising from collapsing bubbles in housing markets, there is an obvious need for us to better understand the evolution dynamics of bubbles in housing.

1.2 Literature Review

Earlier studies on housing bubbles can be classified into three main schools of thought. The first pertains to human psychology [8], where bubbles are treated as manifestations of irrational exuberance in housing markets [9, 10]. Here, bubbles are thought to be created through herding and reinforcing behaviours of market participants. In these studies, the existence and extent of the bubble are typically measured through surveys [11, 12, 13] which are carefully designed to detect such ir-

rational enthusiasm. The second paradigm takes root in economics, where the home prices are intrinsically linked to microeconomic factors such as supply and demand [14, 15], as well as macroeconomics factors such as interest rates [16] and income ratios [15, 17]. Here, models built using empirical data or agent-based simulations [4] are used to relate the economical factors in concern to the intrinsic housing prices. Bubbles are therefore considered phases when prices in the market are significantly more elevated than those predicted by the models [18]. The third approach, which is mainly data and analytics driven, assumes a housing market to be usually stationary, but becomes explosive (non-stationary) as the housing market approaches a bubble. Such non-stationarity can then be tested using time-series analysis techniques, such as the presence of a unit root [19, 20, 21, 22].

The merits of these three different paradigms are obvious, since each emphasize respectively, the psychological, economical and mathematical natures of housing markets. Yet, many of the methods reviewed tend to over-emphasize a single factor, despite giving merits to other approaches. Indeed, each of these paradigms do not contradict but compliment each other and should not be considered in isolation. However, to study the individual factors all at once poses a dimensionality challenge, since housing markets consist of many heterogeneous stakeholders such as buyers, sellers, agents and legislators, with non-linear interactions. To complicate matters, the factors and agents interplay so any single event can spark off a chain of other events. For example, legislators can enact rules (interest rates, economical) which in turn affects the psyche (psychological) of buyers and investors, and so forth. These factors influencing the dynamics of housing markets cross traditional disciplinary boundaries and therefore requires interdisciplinary approaches to tackle [23]. Besides, given the multi-agent heterogeneity and their non-linear (often also non-homogeneous) interactions, many are now recognizing the need to treat housing markets as complex systems [24] and study housing markets with the help of complexity theory.

1.2.1 Econophysics - An Interdisciplinary Approach

The field of ‘Econophysics’ - a term coined by H.E. Stanley during the 2005 Econophysics colloquium in Kolkata, is an interdisciplinary field combining the tools and methods commonly used in the study of physical systems, and applying them on economic systems. Despite this relatively recently term, the relevance of Econophysics date as far back to the 1900s. In the Spring of 1900, Louis Bachelier, a French mathematics PhD student from the University of Paris published his dissertation titled “Théorie de la spéculation” analyzing randomly moving stock prices and their (financial) derivatives. Unfortunately, his work was not well received by other mathematicians then¹, as his work was not traditionally under the domain of Mathematics. It is only five years later, in 1905, when Einstein’s work on Brownian Motion [25] and of stochastic processes caught the attention of the Physics community and highlighted the mathematical importance of Bachelier’s study. While the works of Einstein provided yet another confirmation of the existence of atoms and molecules, one can speculate that these similar equations derived in Bachelier’s case was never meant to describe the collisions and diffusion of solid particles. Nonetheless, it is heartening to know that Bachelier’s thesis, considered by many as the pioneering works in financial mathematics, and Brownian motion, describing the stochastic nature of particle movements, both stood on the same mathematical footing. Fortunately for us in the recent time, many have grown to embrace the importance of interdisciplinary work and recognize the contribution of econophysics [26].

One of the successes of econophysics lies in its ability to simplify complex economical systems to fairly simple physical systems. Relating to the example above, suppose we interpreted the works of Bachelier in the context of Einstein’s findings, one simplistic way to model the price of a stock, is to assume that it performs Brownian motion. In the long time limit, the final position of the particle (or stock, in Bachelier’s case) should be distributed by a normal distribution [27]. However, since the stock price is bounded by $[0, \infty)$ and cannot go below zero value, it is more

¹One of the most famous critics of Bachelier’s thesis was Paul Levy, who would in the future realize his mistakes and reconcile with Bachelier

reasonable to assume that it is the returns of the stock $R = r_1, r_2, \dots, r_n$ over the time period n , that is normally distributed [25]. The return at each time period t , r_t is given by,

$$r_t = \frac{p_t - p_{t-1}}{p_{t-1}}. \quad (1.1)$$

This important insight of econophysics led to the derivation of the Black-Scholes-Merton formula for pricing of financial derivatives, which won the 1997 Nobel Memorial Prize in Economic Sciences. However, during October 2010, a flash crash caused the DJIA to lose almost 9 percent of its original value, only to gain it back within minutes. The probability of observing such an event happening if predicted using a normal distribution, would require the lifetime of the Universe. This event shows the flaws of using a Normal distribution to model returns. The Brownian motion assumption that the behaviour of a particle (or the stock price) was independent of its past is not necessarily true in times of financial crisis! With high volatility, traders tend to trade on impulse due to fear or greed. A better theory was therefore needed to address financial market's apparent ability to show such wild fluctuations presumably due to processes with memory of the past and agent interaction [28]. Investigating this phenomenon, Mantegna and Stanley looked at the return distribution of stocks of different time scales [29]. It can be seen that a fat tail (decays slower than a Gaussian/Normal Distribution) is a universal trait of the return distribution. Such power-law or truncated power-law behaviour have also been independently observed in other studies [28, 30]. These results of this findings provides a better understanding of such wild shocks, and remind us of the importance of distributions in our understanding of the dynamics and behaviour of a particular market.

Although a substantial amount of work relating to determining distributions have been investigated for liquid markets (such as the stock market and foreign exchange markets), a significantly less amount have been conveyed in the less-liquid housing markets. The earlier studies often attribute housing prices to several external factors, such as size, location and other fundamentals. Doing so, hedonic models accounting for such factors have been suggested and used extensively [31, 32, 33] to determine housing price distributions. Unfortunately, few of these models have been compared to empirical data. Besides, these models often do not account, or at most only

take an aggregated approach when considering the interactions between stakeholders (such as buyer, sellers, developers) within housing markets. As with the case of the stock market, these interactions in housing markets can also form herding and positive feedback mechanisms and are usually amplified when housing markets are in a midst of bubbles [34]. As such, due attention needs to be given to interactions at these microscopic levels.

1.2.2 Housing Markets as Complex Systems

Complex systems are generally multi-agent systems, with large numbers of heterogeneous agents interacting with each other, in non-linear and non-homogeneous ways [35, 36]. Due to the non-linear interactions, complex systems exhibit multi-stability, and can be found in different stable regimes. Each regime is made up of an attractor and its basin of attraction. Inside a basin of attraction, the system is stable and robust to small disturbances and perturbations (see Figure 1.1) [37]. However, if the system can only remain within an attractor, its behaviour cannot be considered complex. For a system to be considered complex, it must at some point be able to escape the basin of an attractor, explore other parts of the phase space, before eventually ending up in another attractor.

We think of such transitions between attractors as ‘regime shifts’, which can be understood using bifurcation theory [38, 39]. As the complex system approach a critical transition, the bifurcation parameter approaches a critical value, flattening the potential landscape. Generally, when this happens, the system takes a longer time to equilibrate when perturbed. In some cases, perturbations can also cause the system to be ‘pushed’ out of its basin of attraction, and fall into the basin of another attractor, leading to a regime shift. These regime shifts are rapid, causing the system to experience sudden and catastrophic shocks.

Motivated by these transitions in climatological and ecological systems [39, 40, 41, 42], Scheffer *et al.* called for us to treat these systems as complex systems [43], and listed early warning indicators we can measure to see if the system is indeed approaching a critical point [38]. Suppose we observe the fluctuations of a system approaching a critical transition in response to the gradual changes to the potential

Figure 1.1: A schematic illustrating the approach to a critical transition. [Left] Highly stable state inside a deep potential well, and system will quickly equilibrate after any perturbations. [Mid] Close to a critical transition the potential landscape flattens so the system takes a longer time to recover from perturbations. [Right] At the critical point, the original stable state disappears, and the system experiences a regime shift, going into another stable state.

landscape (See Figure 1.1). Since the system take a longer time to equilibrate as it loses stability, this response time will diverge as we move towards the critical point. We should thus expect the Fourier spectrum of these fluctuations to be shifted and concentrated towards the lower frequencies. This is also known as *spectral reddening*, associated with an increase in the characteristic time scale [44]. From the perspective of the system, its dynamics are perceived to have ‘slowed down’ due to the decrease in recovery rates. If we take snapshots of the system, we can expect it to appear self-similar because its slowed response, especially if the time interval between these snapshots are small. This phenomenon known as *critical slowing down* [45, 46] and causes an increase in autocorrelation, especially at small lag times [38, 42, 39].

It is also common to detect changes in the distribution of these fluctuations. As the system approaches the critical point, the system loses its stability and fluctuations tend to be wilder, which causes an increase in its variance [47]. In addition, the bifurcation may affect the potential landscape in asymmetrical ways. In such cases, we can also expect the distribution of fluctuations to be skewed [48].

The early warning indicators are generic and apply to a large class of regime shifts. Therefore, we can expect housing markets, which are complex systems, to exhibit early warning indicators. Using these indicators, Tan *et al.* detected early

warning indicators preceding to the Subprime Loan Transition (which led to the US subprime housing bubble) as well as the Lehman Brothers Crisis (which led to the Global Financial Crisis) [49]. This findings affirm the early warning indicators as generic tools that can be applied to various different forms of regime shifts, including the analysis of housing bubbles.

1.3 Aim and Objectives

Although a fairly substantial amount of work have been done on the studies of housing bubbles, most of these do not provide a clear definition of a bubble [50, 51]. The reason for this is a lack of understanding of what the ‘normal’, non-bubble housing market looks like. Since every housing market is different and have different rules governing them, are these ‘normals’ universally similar? We aim to address this issue by identifying the distributions observed for housing markets in equilibrium. Next, we look out for deviations to these distributions and design statistical tests which will allow us make a clear distinction between ‘normal’ and bubble. This will provide us with a clearer and unambiguous definition of housing bubbles.

Also, since housing markets are dynamical complex systems, its dynamics as it approaches a bubble would be inherently different from its dynamics near equilibrium. Since the onset of bubbles are usually fairly rapid, amplified by anxiety, herding and positive feedback loops [3], we should expect to find early warning signals associated to these microscopic behaviours. After detecting a bubble, we can confirm the presence of these generic early warning signals, by using indicators that can detect such signals [45, 46]. In this way, we can associate critical slowing down signals to known bubbles.

Next, to gain deeper understanding of housing market dynamics, we must be able to model its behaviour and interactions. To do this, we start by modeling the empirical home price distributions at equilibrium. Following this, we aim to account for the interactions between agents in the model by implementing housing portfolio strategies for these agents. Lastly, we aim to program this as an agent-based model, in order to do *in vivo* examinations of our toy housing market.

1.4 Organization

This thesis consist of a total of five chapters, with this chapter being the first, presented as follows.

In Chapter 2, I present our work on the Singapore housing market, and compare it to the property market in the Greater Taipei Area, using data provided by Chung-I Chou (Chinese Cultural University) and Sai-Ping Li (Academic Sinica). In collaboration with Chung-I Chou, Sai-Ping Li, Shang-You Tee (Nanyang Technological University, currently CP Kelco) and Siew Ann Cheong (Nanyang Technological University), we show the existence of equilibrium distributions in housing markets. Addressing the lack of clear definition for the term ‘bubble’, we provide an unambiguous definition that bubbles are deviations from equilibrium distributions. Treating the equilibrium distribution as a null model, we can attach statistical significances to bubbles. The works is also published as Ref. [51].

In Chapter 3, I present our work on detecting early warning indicators in housing markets, in collaboration with Siew Ann Cheong, K. Y. Michael Wong (Hong Kong University of Science and Technology), Chihiro Shimizu (National University of Singapore) and their students, Marcus Ho (Nanyang Technological University) and Junyi Wei (Hong Kong University of Science and Technology). Using housing data provided by Chihiro Shimizu, we show that generic early warning indicators can be used to detected transitions in housing markets. In particular, we find early warning indicators associated with spectral reddening and critical slowing down in the Singapore condominium market segment. In our comparison study, we also find a possible bubble in Hong Kong (HK) during the period 2006 to 2010, a period of explosive growth in the HK housing market, thereby affirming the potential of generic early warning signals when applied to housing markets.

While considering specific distributions when dealing with housing market, we deal with real world data that is noisy and comes in different sample sizes. This force us to consider the more general question of comparing data of different sample sizes across different distributions. In collaboration with Boon Kin Teh (Nanyang Technological University), Sai Ping Li and Siew Ann Cheong, we quantify the extent

of errors when performing parameter estimation and significance testing on finite samples. Developing correction formulas to account for these effects, we show in Chapter 4 that it is possible to meaningfully compare fits across different distributions for noisy samples of different sizes.

Finally in Chapter 5, I summarize the key findings of the works presented in this thesis and reiterate key contributions of these studies. In particular, I highlight areas where the work presented in this thesis have reinforced our understanding and addressed some of the current gaps in the housing market literature. Lastly, we discuss some of the future research questions and directions, informed by the works presented in this thesis.

Chapter 2

Bubbles are Departures from Equilibrium Housing Price Distributions: Evidence from Singapore and Taiwan

2.1 Introduction

Housing prices and affordability have always been an intensely debated topic because it involves the livelihood of people. For most people, a home is the most expensive asset they will own in their lives. In general, most homeowners take up housing loans, and thereafter commit a significant fraction of their incomes to repay them. Their homes therefore constitutes a large part of their wealth. Hence, housing market crashes affect much broader cross sections of societies than do stock market crashes. In fact, the situation is especially dire for people who bought homes during the height of the bubble. In the most recent case, the Subprime Housing Crisis in the United States (US) was a primer that led to the global financial meltdown in 2008 [7, 9, 52]. To make matters worse, housing markets are illiquid [4], and take a long time to recover [53], holding back the recovery in the rest of the economy.

While the US is still recovering from the Subprime Crisis, a regional housing

bubble of unprecedented scale has been brewing in many Asian cities since 2007. However, because there are no widely accepted definitions for housing bubbles, many Asian governments have denied they are facing such problems [54, 55, 56]. With the economies of US and Europe still struggling to get back onto their feet, a collapse of the Asian housing bubble will send shock waves all over the globe, and throw the global economy back into recession. This is a doomsday scenario no one wishes to come to pass.

Many previous works on detecting bubbles have been done using carefully designed surveys [11, 12], which measures the attitudes of people towards bubble formation using questions, such as those listed in [12]. Another popular and more quantitative method for detecting bubbles include using unit root tests to test for non-stationarity in the fundamental values and asset prices. Using this approach, bubbles have since been discovered in Hong Kong [19], United Kingdom (UK) [20] and several cities in the US [21, 22] for various time periods. More recently, Meng *et al.* computed the correlation matrix of the Housing Price Index (HPI) in 51 US states and found the eigenvalues increasing beyond the maximum predicted by random matrix theory during a period coinciding with the subprime mortgage crisis [57]. Still, others treated housing bubbles as precursors to critical transitions, and have looked for the presence of log-periodic power laws [24, 58, 59], as well as spectral reddening, increasing autocorrelation, and increasing variance as we approach the housing market crash [49]. The latter three are generic early warning indicators associated with critical slowing down as a complex system (which housing markets can be classified into) approaches a regime shift [38]. Critical transitions and early warning indicators in housing markets will be further discussed in Chapter 3 of this thesis.

In discussions with policymakers, we often find it difficult to convince them that the signatures we see in the data points correspond to the presence of a housing bubble, because there are no unambiguous definitions of what a housing bubble is. We therefore believe that it is important to first understand the housing market equilibrium (if it exists) before we can develop an understanding of housing bubbles. How can the prices of real estate be fundamentally determined? Since this

can be attributed to several external factors, hedonic models accounting for such factors have been suggested and used extensively, most notably in the construction of housing indices [31, 32]. While such hedonic models have been compared against traditional pricing models (such as the Dutot model) [33], few have tested these hedonic models against the empirical home price distribution. A recent work by Ohnishi *et al.* assumed that the price of a house P can be treated as a product of a large number of random variable, each corresponding to an external factor F_i . Here, F_i represents the i -th factor, corresponding to size, location, $\dots$, which we can write mathematically as,

$$P(F_1, F_2, F_3, \dots, F_N) = \prod_{j=1}^N F_j. \quad (2.1)$$

Taking the natural logarithm on both sides, we get

$$\log P = \sum_{j=1}^N f_j, \quad (2.2)$$

where we have used the notation $f_j = \log(F_j)$. In the limit of infinitely many factors ($j \rightarrow \infty$), the generalized central limit theorem predicts that the home price distribution will converge (weakly) to a log-normal distribution [60], if we satisfy the Lindeberg condition,

$$\lim_{j \rightarrow \infty} \frac{\max(\sigma_j)}{\sum_j \sigma_j} = 0, \quad (2.3)$$

where σ_j is the standard deviation of the distribution of f_j .

Indeed, many of the early studies [60, 61] on housing typically assume and fit the housing prices to log-normal distributions. However, when fitting to empirical data, Ohnishi *et al.* found that the distribution of housing prices in Japan displayed instead a power-law tail for several years [60]. Similar power-law tails have also been observed in the housing price distribution of London [62]. This happens because the convergence condition (2.3) requires both a large number of random factors ($j \rightarrow \infty$), as well as for the means and variances of f_j to be finite. Practically, the condition is stringent and therefore the asymptotic power laws can be expected should one or more of these means and variances of the distribution f_j be infinite.

In general, it is not clear that a power-law tail in the home prices necessarily indicates the presence of a housing bubble (as in the case of Tokyo). It is important

to first establish what kind of distribution we ought to see when the housing market is in long-run equilibrium. As such, our approach is to look at transaction-level housing data over many years, to identify the equilibrium home price distribution. Once this is known, it should be possible to identify out-of-equilibrium features that might correspond to bubbles. Instead of focusing on the differences between cities, we would like to discover universal features of housing market equilibrium and bubbles that are recognizable in all cities.

Together, Singapore, Taiwan, Hong Kong, and South Korea are known as the Four Asian Tigers. Culturally, Singapore, Taiwan and Hong Kong have a majority of ethnic Chinese and therefore share similar views towards the importance of home ownership [63]. Thus as a proof of concept, we compare the housing markets between Singapore and the Greater Taipei Area (GTA) in Taiwan. In this study, we look across housing type classifications in Singapore and in Taiwan and match the housing types (See Figure 2.1). We find corresponding housing types for the rich of both the GTA and Singapore. In addition, apartments in estates with amenities, such as condominiums in Singapore and Zhu Zai Da Lous in the GTA appeal to young professionals in both cities. Beyond these similarities, the Singapore and Taiwan housing markets diverge, with subsidized public housing in the former and low cost private housing in the latter making up the rest of the market. The detailed comparison between the two housing markets can be found in Appendix A.1.

In Figure 2.1, we show the different housing types in Singapore (SG) and Taiwan (TW). The Singapore housing market is segregated into the landed properties, condominiums and Housing Development Board (HDB) flats. The first two are private properties while the HDB flats are public housing that enjoy subsidies. Taiwan has five housing types - Tou Tian, Zhu Zhai Da Lou, Tao Fang, Hua Sha and Gong Yu. In this report, we compare the landed properties (SG) to the Tou Tians (TW) and the Condominiums (SG) to the Zhu Zhai Da Lou (TW) and Hua Sha/Gong Yu (TW), drawing parallels between the housing types with respect to the type of occupants they appeal to. We also show that the home price distributions in both cities take on a Gibbs-Pareto form, which is well known from the study of income and wealth, as was discussed in Chapter 1 of this thesis. Finally, we report the pres-

	Singapore	Taiwan
↑ Cost / Luxury / Status	<u>Landed Properties</u> <ul style="list-style-type: none"> • Detached • Semi Detached • Terraces 	<ul style="list-style-type: none"> • Tou Tian (透天厝)
	<u>Condominiums</u> <ul style="list-style-type: none"> • Condominiums • Executive Condominiums * 	<ul style="list-style-type: none"> • Zhu Zai Da Lou (住宅大楼) • Gong Yu/Hua Sha (公寓/華厦) • Tao Fang (套房)
Public Housing	<u>Housing Development Board Flats</u> <ul style="list-style-type: none"> • HUDC Apartment • Executive Apartment • 1-5 Room Flat • Studio Apartment 	<ul style="list-style-type: none"> • Social Housing • Public Housing

* Executive Condominiums only become fully private after the first 5 years of minimum occupancy and may be purchased using government grants.

Figure 2.1: A comparison between housing types in Singapore and Taiwan. Shaded in blue are landed properties, which are detached standalone houses, while those shaded in green are apartments/estates with common amenities that can be shared by residents. Public housing is shaded in red.

ence of a housing bubble in the condominium market segment, in Singapore and a emergence of a possible bubble in the GTA.

2.2 Data and Methods

We obtain the Singapore housing price data from two sources. First, the Urban Development Authority (REALIS - Real Estate Information Systems) database captures transactions of private properties dating back to January 1995 [64]. The information provided for each transaction includes the sale price, type of housing, the price per square foot ($\$/\text{ft}^2$) as well as the location of the property. The parcel type (land/strata) and lease type (freehold/99-years/999-years) is also reflected in the data. Second, the HDB data is obtained from the official national statistics website of Singapore [65]. It consists of the town in which the property is located, the flat type (number or rooms), date of transaction, the size of the property as well as the transaction price of the property.

The data for the property transaction price of the GTA can be obtained from the Department of Land Administration, Ministry of the Interior, Republic of China [66], and consists of the transaction data of the price, floor area, type of housing, city and district of the property for the entire Taiwan from 2010 to 2012.

Spatially, we segregate Singapore into 9 different sectors consisting of a total of 28 districts. The map illustrating the different districts and sectors is shown in Figure 2.3(a). The 9 different sectors correspond to the planning regions in Singapore. The 28 districts are the historical postal districts in Singapore, which are often used by real estate agents and representatives in solicitations. The GTA of Taiwan consist of Taipei City, New Taipei City, Keelung City and Taoyuan City, which forms the Greater Taipei-Keelung-Taoyuan metropolitan area.

For a meaningful comparison of the data across different years, all home prices are discounted (See Appendix B) using the consumer price index (CPI) of Singapore/Taiwan. The historical CPI of Singapore is obtained from the Singapore Statistics website [67], while the historical CPI of Taiwan is obtained from National Statistics, Republic of China [68]. Since the CPI is reported monthly, we use the simple average (total for each month divided by twelve months) to compute each year's CPI. The CPI is then used to discount the price to the base year, 2014 (Singapore) and 2011 (Taiwan). We have chosen the CPI over other indicators as a basis for discounting because it reflects the overall spending power of an individual/household relative to the base year.

The CDFs of respective data sets are then fitted using both an exponential and a Pareto distribution. In the exponential fit, the per square foot (PSF) price, x , is fitted using the distribution,

$$P_{exp}(X = x) \propto \exp\left(-\frac{x}{T}\right), \quad (2.4)$$

where T is the temperature of the Boltzmann-Gibbs distribution, $T = 1/\beta$. In the Pareto (Power-Law) fit, we use,

$$P_{PL}(X = x) \propto x^{-\alpha} \quad (2.5)$$

where α is the the Pareto exponent. Each of the fits are performed using the Clauset-Shalizi-Newman (CSN) method [69], which we will briefly describe here, but

elaborate upon in Chapter 4. For each data set, we use the Maximum Likelihood Estimation (MLE) method to estimate the ‘temperature’, T of the exponential distribution, the exponent α of the Pareto distribution and the threshold x_{min} above which this distribution holds. Next, using a re-sampling technique, we run p -tests to check the statistical significance of the fits to (a) a Pareto distribution, and (b) an exponential distribution. A p -test value of $p > 0.05$, which is half the value suggested by CSN is used as an acceptance criterion. We also visually inspect the plots to identify the presence of any significant outliers, which could possibly correspond to bubbles. These include positive deviations to empirically fitted data (data points significantly above the fitted distribution) and upwards flexing of the tail (decrease in the value of α).

Significant positive deviations from an empirically fitted distribution as detected by observation are indications of possible *Dragon Kings* (DKs). DKs are a result of positive feedback mechanisms and are systematic deviations. To test for the presence of DKs, we use the U-test proposed by Pisarenko and Sornette [70], but with modification for the use on the exponential distribution (2.4). To do this, we ordered the empirical data set in descending $X = \{x_1, x_2, \dots, x_N\}$ and obtain the x_{min} and parameter $\hat{T}$ using the MLE procedure. Using the estimated parameters x_{min} and $\hat{T}$ we obtained the estimated CDF, $F(x|\hat{T})$. By inserting the set X back into its CDF, we will obtain a set U which is distributed by the standard uniform distribution, where $0 < u < 1$, assuming the exponential distribution is a good estimate of the empirical data set X . This is known as the probability integral transform or the inverse CDF transform (see Appendix C.1). The k -th ranked sample of the set U is distributed with probability (see Appendix C.2),

$$f_k(u) = \frac{n!}{(n-1)!(k-1)!} u^{k-1} (1-u)^{n-k}. \quad (2.6)$$

The probability that the observed x_k is an overestimate of the true distribution is (from the empirical observation), or p -value, can be given by a sum of all possible values from the empirically fitted values to the maximum value, which equals to

unity. In the continuous limit, we have,

$$p(x_k) = \frac{n!}{(n-k)!(k-1)!} \int_{Z_k}^1 u^{k-1} (1-u)^{n-k} du = 1 - \text{betainc}(Z_k, n-k+1, k), \quad (2.7)$$

where $Z_k = F(x_k)$ is the empirical observed cumulative probability of the k -th ordered term, *betainc* is the normalized incomplete Beta function and $F(x_k)$ obtained using the empirically fitted parameters. This measures the probability of the empirically observed sample deviating from its fitted distribution. We follow Pisarenko and Sornette by classifying p -values with $p < 0.10$ or $p < 0.90$ as DKs [70].

Finally, we locate the housing transactions corresponding to the outliers. Using this information, we analyze the similarity between anomalous transactions spatially and temporally. We present this information in the form of a heat map of housing prices in Singapore. For each sector, we obtained the 75th-percentile price of all the housing transactions in a month. The 75th-percentile prices for each region and time frame is color coded, with red signifying prices closer to the maximum observed value throughout the years and blue signifying prices closer to the minimum observed value throughout the years. By observing the price change in both space and in time, we are able to clearly see the emergence and evolution of the bubble.

2.3 Results

2.3.1 Equilibrium Home Prices

We start by reporting our findings on the Singapore housing market. Figures 2.2(a), 2.2(b) and 2.2(c) show the cumulative distribution functions (CDF) of price per square foot (PSF) for HDB apartments, condominium apartments, and landed properties. For the condominium and landed properties the data we fitted was from the period 1995 to 2014, whereas the HDB properties data was from 2000 to 2014. We fit part or all of these CDFs to the exponential (Gibbs), power-law (Pareto) distributions, and found that the CDFs of HDB and condominium apartments are best fitted to decaying exponentials (with a log-log plot inset showing a poor fit to the power law). On the other hand, the tail of the landed properties' CDF is best fitted to a power law (with a semi-log inset, showing poor fit to the exponential distribu-

tion). The results of all fits and tests of statistical significance for this chapter are listed in Table 2.1.

Table 2.1: The p -test values for housing and income distributions in Singapore and Taiwan. Due to the presences of Dragon Kings (DKs) in the the distribution of the condominiums, Da Lou and Gong Yu/Hua Sha, the p -test is not conclusive for these distributions.

Data	Distribution	p -value	Remarks
Singapore HDB	Double Exponential	0.10	
Singapore Condominiums	Exponential	-	DKs
Singapore Landed	Pareto	0.62	
Taiwan Toutian	Pareto	0.02	
Taiwan Da Lou	Pareto	-	DKs
Taiwan Hua Sha/Gong Yu	Pareto	-	DKs
Singapore Wealth (Forbes)	Pareto	0.200	Finite Size
Taiwan Wealth (Forbes)	Pareto	0.025	Finite Size
Taiwan Income	Gibbs-Pareto	0.513	

In Appendix A.2 we show the contributions to the CDFs from the various years for the HDBs, Condominiums and Landed Properties respectively. For HDB apartments, the yearly contributions to the CDF remain largely the same from 2000 to 2009. This is also the case for condominium apartments and landed properties for the period 1995 to 2006. This suggests that the HDB and condominium segments were in stable equilibria over these periods, and their housing market equilibria are characterized by exponentially decaying CDFs. For the landed properties, the equilibrium distribution is a power law.

Figure 2.2: The per-square-foot (PSF) home price CDFs of (a) the HDB flats, (b) condominium apartments and (c) landed properties. The CDFs are plotted in a semilog, semilog and loglog scales respectively. For each of the sub-figure, the alternate distribution is shown in the inset. (a) The HDB flats follow a double exponential distribution with temperatures $T_{H1} = \$58.4/\text{ft}^2$ and $T_{H2} = \$38.4/\text{ft}^2$. The second exponential is formed by transactions after 2009. (b) Condominium apartments PSF roughly follows an exponential distribution with temperature $T_C = \$493/\text{ft}^2$, except for a bump between $\$3000/\text{ft}^2$ and $\$4500/\text{ft}^2$ which corresponds to a bubble that started in 2007. (c) The CDF of the landed properties follows a power-law distribution with Pareto exponent $\alpha_L = 6.24$. For all three sub-figures, the different colors represent a transaction in one of the sectors shown in Figure 2.3(a)

Figure 2.3: The district/sector map with the upper quartile price of the different property types in Singapore. (a) Map of Singapore with the geographical location of the 28 postal districts (in numbers), segregated into 9 sectors represented by their respective colors. Plots of the yearly upper quartile price for (b) the HDB flats, (c) the condominium apartments and (d) the landed properties, for the 9 sectors in Singapore.

2.3.2 Departures from Equilibrium as Housing Bubbles

In Figure 2.2(b), we see that the condominiums data generally fits well to an exponential distribution with ‘temperature’ $T_C = \$486/\text{ft}^2$, relating to the mean of the upper quartile price distribution, but notice a bump that stretches from $\$3000/\text{ft}^2$ to $\$4500/\text{ft}^2$ ¹. The year-by-year condominium price distributions from 1995 to 2014 are shown in Appendix A.2. From these plots, it is evident that the bump first appeared in 2007, and persisted till 2014. This is in agreement with the upper quartile price spike we see in Figure 2.3(c). This bump is a large positive deviation from the Gibbs distribution, which suggest a higher-than-expected probability of finding a property selling in that range, than expected from an equilibrium exponential distribution. Didier Sornette calls such statistical features *Dragon Kings* (DKs) [71]. To confirm this, we applied the U-test [70] to the condominium CDF using the exponential distribution as the reference. The results of our U-test suggest that there is at least one DK for prices beyond $\$3000/\text{ft}^2$, which is the beginning of the bump. This observation provides us with a technical and unambiguous definition of a housing bubble, i.e. a dragon-king out-of-equilibrium feature in the home price CDF, and also a statistically rigorous way to test for the presence of the bubble. More importantly, the price distributions of the HDB and landed properties show no positive deviations from their respective equilibrium distributions, which suggest that there are no bubbles in these two market segments.

From the color scheme in Figure 2.2 we see that the transactions in the bump and beyond are located in the prestigious districts 9 and 10 which make up Sector 2 (magenta) in Singapore (see Figure 2.3(a)). Situated near the city center, we suspect these properties are ‘investment class’ and susceptible to speculation. To confirm this, we plot the yearly upper quartile price of each sector (represented by their respective colors in Figure 2.3(a)) in Figure 2.3(b). We have chosen the 75th-percentile price instead of the mean as we are interested in the tail end of the distribution but at the same time do not want our analysis to be affected by outliers. Prior to the bubble, the movement of prices in all sectors are generally in tandem,

¹All units in \$ refer to the Singapore Dollars, while all units in NT\$ refers to the New Taiwan Dollars

as in the case of the HDB and landed properties segments. Following the start of the bubble in 2007, a sudden spike (in 2007) and dip (in 2008) in the condominium upper quartile prices can be seen in several sectors. In particular, the dip which we suspect was in response to the 2008 Global Financial Crisis is deepest for the prime areas (Sectors 1 and 2). Given that this dip is weaker in sectors further from the prime districts, and in some sectors the prices actually rose, this suggests that transactions near or within the prime areas were of a speculative nature, rather than for owner occupation. A repeated sales regression model done by Jiang *et al.* also detected an explosive period from early 2007 to early 2008 [72]. We further verify this explosive growth in the next subsection.

2.3.3 Spatial-Temporal Dynamics of the Singapore Housing Market

To visualize the spatial and temporal dynamics simultaneously, we constructed a monthly sequence of heat maps (See Appendix A.3 to observe how the increase in prices gradually spread across Singapore. We see that the home prices for the condominium apartments first started to rise sharply in the prime districts 9 and 10 (Sector 2, magenta), before spreading into the neighboring sectors. This spiking phenomenon is especially clear at the height of the housing bubble in September/October 2007. Then as the bubble slowly relaxed in the prime districts in 2008 and 2009, home price increases continued to spread outwards into districts further away. By the end of 2009, the upper quartile in each sector has risen compared to earlier years for the whole of Singapore. This strongly supports the suspicion that the bubble first formed in districts 9 and 10, possibly from speculative behavior, around 2007. Such spreading effects arising from Sector 2 are also observed in the heat maps of the Landed and HDB segments, but the pulsatile dynamics are not observed in both the HDB flats and the landed properties, where there are no bubbles.

The spatial and temporal dynamics for all three housing types show that price increase generally starts from the prime districts before spreading out to the periph-

eral districts. Across housing types, we see that the price increase was led by the condominiums which peaked in 2007, followed by the peaks of the HDB (2011) and landed properties (2010-2012). We also found a second exponential part in the HDB price distribution appearing after 2009, after the bubble formed in the condominiums. Therefore, we suggest that these purchases are likely to have been made by those who are affected by the condominium housing bubble. However, more data will be needed to better understand this phenomenon.

2.3.4 Comparison between Singapore and Taiwan Housing Markets

Figure 2.4 shows the fits to the PSF CDFs of the different housing types in the GTA of Taiwan. We see that the Tou Tian is Pareto distributed ($\alpha_{TT} = 2.81$), and we can see this as a straight line in the loglog plot. Both the Zhu Zai Da Lou and Gong Yu/Hua Sha follow exponential distributions for most part of their bodies, but deviates positively at their tails.

Figure 2.4: The PSF CDFs of the different property types in the GTA. The CDFs of the (a) Tou Tian (b) Da Lou and (c) Gong Yu and Hua Sha, shown in loglog, semilog and semilog scales respectively. For each plot, the alternative distribution is shown as the inset. We see that (a) the Tou Tian is power-law distributed with a Pareto exponent of $\alpha_{TT} = 2.81$, (b) the Da Lou in the GTA largely follows an exponential distribution with $T_{DL} = \text{NT}\$1.25 \times 10^5$ and (c) the Gong Yu/Hua Sha in the GTA follow an exponential distribution with $T_{HSGY} = \text{NT}\$1.17 \times 10^5$. For (b) the Da Lou and (c) Gong Yu/Hua Sha, we observe large positive deviations in the tail of the distribution that indicates the presence of a bubble.

The Tou Tian are detached houses similar to landed properties in Singapore. Ownership of such houses is symbolic of one’s status and hence these houses are usually owned by the rich in the respective cities. The Zhu Zhai Da Lou and Tao Fangs are very similar to the condominiums and shoe box condominiums in Singapore. These homes usually have common amenities such as swimming pools and other recreational facilities, appealing to the upper-middle class. The Gong Yu and Hua Sha which resembles more like the HDB flats in Singapore follows an exponential distribution.

As we can see, the equilibrium distributions of corresponding housing types are the same between the Singapore housing market and Taiwan housing market. We observe that the more expansive and desired housing type is Pareto distributed in both Singapore and in the GTA. The upper-middle class housing types are exponentially distributed, with a possible DK in the Singapore condominiums and in both the GTA’s Zhu Zhai Da Lou and Gong Yu/Hua Sha. This comparative study between Singapore and Taiwan reminds us of a recent study conducted on the Housing Price Index (HPI) of the top ten key cities of China [73]. It is discovered the top tier cities, Beijing, Guangzhou and Shenzhen, collectively referred to as a club, are key drivers in the rise of the HPI index in China. Similarly, we can expect the cities in the GTA to be the driver in Taiwan home prices. However, here in this study, we used raw transaction prices as compared to aggregated HPIs of different cities in the GTA.

2.4 Discussion

2.4.1 Income, Wealth, and Home Price

Singapore tightly regulates the ownership of properties by foreigners. Landed properties can only be bought by Singaporeans and Permanent Residents (PRs) as these are “considered a special class of property that Singaporeans aspire to own” ([74], para. 30). On the other hand, there are no restrictions on foreign ownership of condominiums. However, in recent years rules have been put in place to strongly discourage speculative buying. Finally, the heavily-subsidized HDB flats can only

be owned by Singaporeans and eligible PRs. Regulations governing ownership of specific housing types in Singapore can be found in Appendix A.1. Therefore, there should be no surprise that in Singapore, the rich and higher-income group tend to live in landed properties, while the middle class lives in condominium apartments and HDB flats. Though the housing market is officially classified into three separate segments (see Figure 2.3(a)), this Gibbs-Pareto distribution of home prices mirrors the Gibbs-Pareto distribution, which is a well-known and commonly accepted distribution for income/wealth, consisting of an exponential (Gibbs) distributed body and a power-law (Pareto) tail. This exponential body-power-law tail structure of income/wealth distribution has been verified by data for US and UK [75, 76] and Japan [77], following naturally from a model of wealth produced by additive and multiplicative processes [78, 79].

At present, there are no mechanistic arguments nor are there theoretical arguments suggesting a connection between home prices and wealth. However, since housing is a significant component of a household's wealth, such a close relationship is naturally expected, but not yet verified in real-world data. Going through London's home price data, MacKay was one of the first to suspect a connection between the two distributions [62]. We also suspect such a relationship in Singapore. However, this is tricky to test as the income and wealth data in Singapore is not publicly available. This is because they are collected under the Statistics Act and thus confidentiality is protected under the act. Therefore, we use as proxy the Forbes top 50 wealthiest Singaporeans [80]. We fit the CDF of the Forbes top 50 richest Singaporeans and estimated the Pareto exponent of wealth in Singapore to be $\alpha \approx 2.2$ (see Figure 2.5)

Using income distribution of Taiwan obtained from [81, 82], we plotted the CDF in Figure 2.6. In the main figure, we see that asymptotically, the distribution is a straight line and follows a power-law distribution. In the semilog inset, it is a straight line up to x_{min} which also show that it has an exponential body. However, the exponent for the power-law regime of the Taiwanese income distribution is much higher than those observed in the other parts of the world. For the case of Taiwan, we found that the income distribution for the whole of Taiwan follows a Gibbs-Pareto

Figure 2.5: The wealth distribution of the top 50 richest person in (left) Singapore and (right) Taiwan. The p -value of the fit is given in Table 2.1

distribution with a Pareto exponent of $\alpha = 4.97$ (Figure 2.6), while the wealth distribution for Taiwan, as estimated from the Forbes top 50 wealthiest Taiwanese list [83] has Pareto exponent $\alpha \approx 2.5$.

Comparing the housing price distribution of Singapore and Taiwan respectively to its wealth and income distributions, we see that the Pareto exponent $\alpha_{TT} = 2.82$ of the Tou Tian PSF CDF in the GTA is comparable to the Pareto wealth exponent of $\alpha \approx 2.5$. The small discrepancy can be explained if we take into account the fact that much of the wealth in Taiwan is concentrated in the GTA, and hence we expect the wealth distribution within the GTA to have a fatter tail. On the other hand, the Pareto exponent $\alpha_L = 6.24$ for landed properties in Singapore is significantly different from the Pareto wealth exponent of $\alpha \approx 2.2$. More importantly, we understand from previous studies that wealth distributions generally have power law tails with α a very stable range between 1.5-2.5 [131]. Suppose if everyone owns only one house, and spends the same proportion of his/her wealth on their house, we would then expect the exponents in housing and wealth to be similar. This is however not the case in Singapore. We suspect this is due to a combination of Singapore being a small market and tight regulatory oversight from the government, and have plans to conduct a follow-up study linking wealth and home prices.

Figure 2.6: The CDF of the Taiwan wealth distribution on a log-log plot following a straight line with $\alpha = 4.97$. In the inset, we show the semi-log plot of the same data plotted up to x_{min} . We see that the body of the distribution generally follows an exponential distribution which is seen as a straight line on the semi-log plot.

2.4.2 Bubble and Bubble-Like Features in Home Price CDF

We find three ways the home price distribution can change with time. First, even after discounting the housing price with the CPI, we see the tail of the price distribution is extending towards higher price levels. However, this tail extension is either along the equilibrium distribution, or falling slightly below the equilibrium distribution. This can be seen as a general upward trend in the upper quartile price for all the HDB, condominiums and landed properties in Singapore (Figure 2.3). We do not know whether this also happens for the GTA, since we only have two years worth of data. However, since there is no systematic positive deviations from the equilibrium distribution, we do not consider such cases as bubbles.

Second, we observe positive deviations to the equilibrium distribution, as seen in the condominiums (SG), the Zhu Zhai Da Lous (TW) and Gong Yu/Hua Sha (TW). These deviations are possible DKs which might correspond to a bubble. While this can also be seen as a sharp rise in the upper quartile price, in such plots it is difficult to distinguish DKs from a simple tail extension. In such cases, a suitable statistical test such as the U-test [70] must be carried out.

Finally we may also find an upward flexing of the power-law tail in the housing price distribution. Most power laws in physical systems are known to be robust. For example, the exponent in the Gutenberg-Richter's (GR) Law for earthquakes is always very close to 1, no matter which seismically active region we are talking about. However, Ohnishi et al. [60] observed changes in the power law exponents during the years when Japan was experiencing a property bubble. From our preliminary study of the London housing market, we also observe the same upward flexing of the flats/apartments distribution in London from the period 2012 to 2014. We consider such cases as bubbles, as these case may be a more extreme case of the second phenomenon (positive deviations to empirically fitted distribution) elaborated in the previous paragraph. In fact, upward flexing of the GR Law have been found to precede large earthquakes [84].

2.4.3 Cause of the Singapore Housing Bubble

Just like the US housing bubble was the result of credit easing encouraged by the government, regulatory easing in Singapore is also the likely cause of the bubble formation in the condominiums but not in the other segments. HDB properties and landed properties remain highly regulated, making speculative investments difficult. After the explosive growth in 2007, the Singapore government have intervened in the housing market by imposing a series of cooling measures to discourage speculation and fast turnovers of properties. Had such restrictions been waived as well, these two segments can potentially participate in the housing bubble as well, as demonstrated by landed properties on Sentosa, a small island south of the main island of Singapore, where large price swings have been observed. In a sense, regulatory oversights successfully prevented a housing bubble in the HDB and landed housing market

segments.

2.5 Conclusion

In conclusion, the works presented in this chapter provides a background understanding of housing markets both in Singapore and in the GTA. We find that in both these cities, the equilibrium statistical distribution of home prices follows a Gibbs-Pareto distribution. With this knowledge, we propose a clear and mathematical definition of a housing bubble, which are outliers in the price distribution, and show how to test for the presence of such features. Since the outlier test (U-test) we propose here can be interpreted statistically, we are now able to attach statistical confidence when using it for evaluating housing bubbles, which is never done before in the housing market literature. Using this tool, we find bubbles in the Condominium segment of the Singapore housing market and possible emergence of bubble in the Zhu Zhai Da Lous of the GTA. To understand the emergence of this bubble in Singapore, we plot a heat map of prices which enable us to visualize both the spatial and temporal evolution of home prices in Singapore. We find that the bubble years of 2007 started from the prestigious postal district 9 and 10, eventually pulsating and spreading out to the other peripheral districts. This provides valuable insights to policymakers pertaining to the identification and intervention of housing bubbles.

Chapter 3

Early Warning Indicators in Housing Markets

3.1 Introduction

In the introduction section of Chapter 2, we explained how housing market bubbles and crashes can have long-term detrimental effects on a large cross section of societies. Motivated by the lack of clear and quantitative definition of a housing bubble, we developed and illustrated a simple and quantitative method to identify bubbles in Chapter 2. By identifying the equilibrium distributions in housing markets and interpreting bubbles as departures from these equilibrium distributions, we can attach statistical significances when testing for these bubbles [51]. During this process, we also observed the spatial and temporal features that accompany the formation of a bubble, such as the pulsating dynamics and the formation of out-of-equilibrium features such as a growing bump and fatter tails in the home price distribution. Although such observations can provide some understanding on the spatial and temporal dynamics associated with the formation of a housing bubble, they do not provide any clues to the underlying mechanics and interactions needed for the transition from equilibrium phase to a bubble.

Bubbles can hardly be considered a new phenomenon, with documented history going as far back to the Tulip Mania which popped in 1637 [85, 86]. However, in recent years, use of the term ‘bubble’ becomes increasingly frequent [87], and is

used to describe any asset class whose price is rising rapidly, and possibly super-exponentially. Given the attention we now have on bubbles, it is no surprise that a large number of works has already been done on this subject. In the introduction section of this thesis, we reviewed the literature on the earliest works pertaining specifically to housing market bubbles, and categorize them into three areas of focus. Firstly, works that survey the psychological and behavioural aspects of agents, quantifying and associating the irrational exuberance in agents to housing bubbles [8, 9, 10, 11, 12, 13]. Next are the studies that address the economical aspects of bubbles, which treat bubbles as deviations of housing prices from economic fundamentals [14, 15, 16, 15, 17]. Thirdly, the data driven approaches involving unit root test and/or other variants to detect non-stationarity in time series of housing markets [19, 20, 21, 22].

The works in these three areas have been successful in attributing specific housing bubbles to its causes, but also highlights the broad nature of housing bubbles. In view of this, a general understanding of housing market bubbles cannot be achieved simply by taking specific approaches that tackle only specific types of bubbles in isolation. Since housing markets consist of many heterogeneous, interacting agents (buyers, sellers, developers, etc.) that generally do not share similar goals, these agents often also interact non-linearly. These characteristics have led many to realize and treat housing markets as complex systems, and study them using generic approaches that apply to a more general class of such complex systems.

One of the earliest successes of this approach was presented in a series of works by Sornette *et al.*, on forecasting housing market crashes using the log-periodic power-law singular (LPPLS) model [88]. In the LPPLS model, crashes are modeled as singularities, which are commonly found in non-linear complex systems. Doing so, Sornette *et al.* studied housing market data for the different states in the US [58], the United Kingdom (UK) [24] and Switzerland [59]. Using the LPPLS model, Sornette and Zhou [58] successfully forecast the burst of the US housing bubble [89].

Meanwhile, the successes of Scheffer's works on detecting early warning signals in ecological systems [40, 41, 90] led to the finding of similar early warning signals in other different complex systems [91, 92, 93]. Using only aggregated US housing

market data, Tan and Cheong found generic Schefferian early warning signals, such as increased variance and autocorrelation as well as spectral reddening, associated with the Subprime Loan Transition and Lehman Brother crash in the US [49].

Both the successes of the LPPLS model and Schefferian early warning signals demonstrates the potential of generic indicators in forecasting of bubbles and crashes, by examining the temporal signatures in housing markets. As noted in the spreading of the Singapore condominium bubble from central to peripheral districts in Chapter 2, we can also explore spatial signatures [94] of bubbles. Meng *et al.* computed the eigenvalues of the correlation matrix of the Housing Price Index (HPI) of different US states for different time periods and showed the largest eigenvalue, used typically to represent the collective behavior in markets, experienced a sharp increase in 2008 at the height of the financial bubble [57]. Since each state is geographically distinct, this result provides insights on the spatial dependence of housing bubbles.

Most of the pioneering works in housing market-complex systems described above have focused mainly on the more mature western housing markets. In the recent years however, home prices in many parts of Asia have also risen to unprecedented heights, prompting waves of concerns about an impending crash in this region [55, 54, 56, 95, 96, 97]. Using a similar method highlighted in [57], Meng *et al.* studied the correlation of housing prices in China's 10 key cities and showed that the Chinese housing markets can be segregated into clubs of cities. They also find the average home price in China is driven predominantly by the club consisting Beijing, Shenzhen and Guangzhou [73]. This demonstrates the spatial and demographic dependence of the Chinese Housing bubble.

In close proximity to Guangzhou, and sharing borders with Shenzhen, the Hong Kong (HK) housing market is also a target for Chinese property investors, and is presently experience a housing bubble of unprecedented proportions [98, 99]. Despite this, much of the capital arising from the economic growth of mainland China continues to find it way into the HK housing market. This unique dependence on China adds another layer of complexity to dynamics of the HK housing market.

Collectively, HK, Singapore, Taiwan and South Korea are known as the four

economic tigers of Asia, with strong economic growth during the 1990s. In this study, we focus on two of the four tigers, analyzing the HK residential housing market from the period of 1992 to 2010 and the Singapore condominium market segment from 1995 to 2014. Using tools developed to detect regime shifts in complex systems, we demonstrated that early warning signals can be detected in both housing markets. For the case of the Singapore housing market, the early warning signals coincides with two known economic events, first the Asian financial crisis in 1997, and the second a known bubble period in the Singapore condominium segment which started in 2007, as presented in Chapter 2. We use this as a proof of concept to show a bubble in the HK residential housing market from the period 2006 to 2010.

We organize this chapter as follows. In this introduction, we motivate the study of housing markets as complex systems and show the successes of previous works in separate housing markets. In the Data and Methods section (Section 3.2) that follows, we present both the Singapore and HK housing data used in this chapter, with regards to its acquisition and pre-processing. We also present in the same section, the methods used to detect the critical slowing down indicators used in this chapter. Thereafter, we report and discuss the findings on both Singapore and Hong Kong housing markets and make comparisons between them in the Results and Discussion section (Section 3.3). Lastly, we conclude by highlighting the potential future works enabled through this study.

3.2 Data and Methods

The HK property transaction data is obtained from EPRC Limited, a proprietary database providing property related data warehousing services in HK. The HK dataset contains a sample of the properties transacted in HK from the period 1992 to 2010, which include a total of 5,090,732 transactions for the whole of Hong Kong. Included for each transaction is the date of transaction, the size of the unit (both in gross as well as saleable area), the unit price of transaction in HK\$/ft² (both in gross as well as saleable area), the type of property (eg. Residential, Commercial, Industrial, etc.), the district and the address of the property. Using the address of

the property, we split the HK dataset into the different administrative districts.¹ Filtering the data to include only the residential property transactions, we obtained 2,526,384 records of the housing transaction price data for HK. Next, we discounted the per-square-foot saleable unit housing prices using the HK consumer price index (found in Appendix B), following similar rationale listed in Chapter 2 to obtain the discounted transaction data for HK. The Singapore condominium data which was previously processed as described in Chapter 2 was also used as a comparison in this study.

Since transactions are time-stamped, we obtain the monthly mean price time series $P(t)$, by computing the mean transaction price for each month. Similarly, by counting the number of transactions in a particular month, we are able to determine the volume time series $V(t)$, for both the HK residential data and the Singapore condominium data.

3.2.1 Detrended Fluctuation Analysis

A rising trend in the mean price need not indicate the presence of a bubble. Also, there might be rises and falls due to seasonality in home prices. Therefore, it is important to remove any trends that can be inherently present in the time series before attempting to perform any time series analysis. This is done so that any signals associated with critical transitions are not masked out or polluted by the underlying trend. To do this, we employ Empirical Mode Decomposition (EMD) [100, 101] to decompose the time series into a series of intrinsic mode functions (IMFs). Given a time-series $S(t)$, the EMD seek to decompose S into its IMFs $c_i(t)$ such that,

$$S(t) = \sum_{i=1}^N c_i(t) + r(t), \quad (3.1)$$

and $r(t)$ is a residue function.

The IMFs $c_i(t)$ can be determined using an iterative algorithm (see Figure 3.1) which can be understood as follows. (1) Determine all local maxima and minima

¹Spatially, HK is segregated into 18 administrative districts, are found listed in <http://www.gohk.gov.hk/text/eng/welcome/index.html>

Figure 3.1: The procedure to perform empirical mode decomposition. (a) The original time series. (b) An envelope of maxima and minima of the original time series is plotted in red, and the mean of the two envelopes plotted in black. (c) The mean line in black is subtracted from the original data in blue leaving the green line. Treating the green line as signal, process (a) to (c) is repeated until the threshold in Equation (3.2) is reached. (d) The first and fastest IMF, $c_1(t)$ is obtained and shown in magenta. This is then removed from the original data in (a), then the sifting process from (a)-(d) is repeated to obtain the subsequent IMFs.

in S . (2) Fitting cubic spline to the maxima and minima such that the spline of maxima forms the upper envelope and the spline of minima forms the lower envelope of the time series. (3) Define the mean of the two envelope as $m_1(t)$ (see Figure 3.1(b)). (4) Subtract the mean from $S(t)$ to obtain $h_1 = S(t) - m_1(t)$ (Figure 3.1(c)). (5) Perform steps (1)-(4) on h_1 to obtain $h_{1,1} = h_1 - m_{1,1}$, where $m_{1,1}$ is the mean of the envelopes in h_1 . (6) Repeat step (5) for k iterations to obtain $h_{1,k}$ until the predefined stoppage criterion,

$$\sum_t^T \frac{|h_{1,k-1}(t) - h_{1,k}(t)|^2}{h_{1,k-1}^2(t)} < 0.3 \quad (3.2)$$

is reached (Figure 3.1(d)). The threshold 0.3 is a typical value chosen for the stoppage criterion [100, 102] and T is the length of the time series $S(t)$. Then, $h_{1,k}(t)$ is the first IMF, $c_1(t)$.

Subtracting $c_1(t)$ from $S(t)$ leaves a time series containing the other IMFs. The process (1)-(6) is then repeated to obtain the other c_i 's, shown in Figure 3.2. When all IMFs are found, a residue $r(t)$ containing at most one extremum is left. Then, by subtracting the residue $r(t)$ from the original time series $S(t)$, we get a time series containing only the superposition of intrinsic mode functions, $S(t) - r(t) = \sum_{i=1}^N c_i(t)$.

In Figure 3.2 we see each of the IMFs have varying time scales, with the first IMFs having the shortest period and the residue having the slowest period. As such, the residual mode can be treated as a trend, which varies slowly and monotonically over long time frames.

Using this, we performed the EMD on the price-time and volume-time series for both HK housing data and the Singapore condominium data. The residual mode which contains the trend is subtracted from the respective time-series to obtain the detrended time series, which contains largely the fluctuations in the respective time series. We have chosen EMD technique to detrend the time series over other box detrending techniques (such as Gaussain kernel, moving averages and residual time series) [103] since the IMFs obtained can provide valuable insights to the different time scales and lead-lag relationships (through Hilbert transforms [104]) present in housing markets. This is however beyond the scope of this chapter, but can be used

Figure 3.2: The nine IMFs of the original time series in 3.1(a) with different time scales from $c_1(t)$ (fastest) to $r(t)$ (slowest).

as a basis for future studies.

After obtaining the detrended time series, we make snapshots for each of the detrended time series, $S(t) - r(t)$, by using a 24-months window and sliding each time by a month. The window size of 24 months was chosen so that the annual seasonal effects of home purchases is properly averaged over in each window. For each window with time series $W^{(j)}$, we computed and analyzed the power spectrum and performed autocorrelation analysis to identify regime changes.

In begin chapter of this thesis (see sub-section 1.2.2), we illustrated how the dynamics of systems approaching critical transitions can be studied using detrended fluctuation analysis. Since a system approaching a critical point is gradually losing its stability, the response time of these systems diverges. Suppose we look at the time-series of such system approaching criticality, we can expect a shift of power towards lower frequencies. To detect this (red)-shift, we first used the fast Fourier transform algorithm implemented in *MATLAB* to apply discrete Fourier transform (DFT) on the discretized time series windows $W^{(j)}(t_k)$ for each $k = 1, 2, \dots, N - 1$.

This produces the Fourier spectrum,

$$\tilde{W}^{(j)}(\omega_p) = \frac{1}{N} \sum_{k=1}^N W^{(j)}(t_k) \exp(i\omega_p t_k), \quad (3.3)$$

where $\omega_p = (2\pi p)/T$ are the frequencies with components $p = 0, 1, \dots, N-1$ for the discretized time series. The power spectrum $P^{(j)} = |\tilde{W}^{(j)}(\omega_p)|^2$ can then be obtained using by mod-squaring the result in (3.3). Since we use a 24 months window size, we expect to see 24 frequency bands with 11 repeated and symmetrical frequency with a zero frequency (which corresponds to the DC input in signal processing). Next, we remove the 11 duplicated bands and also the DC signal, leaving us with 12 frequencies bands.

Next, we define the frequency at which the cumulative sum of the truncated spectrum exceeds half the value of its maximum as,

$$\Omega_{1/2} = \min(\Omega \in \{0, 1, \dots, 11\}), \quad s.t. \quad \frac{\sum_{j=1}^{\Omega} P^{(j)}}{\sum_{j=1}^{11} P^{(j)}} > 0.5. \quad (3.4)$$

Since we find at least one diverging time scale as the system approaches a critical transition, we can expect the power spectrum to be concentrated at the low frequencies and hence be red-shifted in these instances. We then look for instances of such spectral reddening, when the value of $\Omega_{1/2}$ is decreased to a low value. We present these results in Section 3.3.

In addition to the $\Omega_{1/2}$ indicator which measures how red-shifted a time-series window is, we also introduce a second indicator, the Pearson autocorrelation ρ . Setting $X = W^{(j)}$ for each window, we also calculate the lag-1 Pearson autocorrelation for each window $W^{(j)}$,

$$\rho = \frac{\sum_{i=1}^{n-1} (x_i - \bar{x})(x_{i-1} - \bar{x})}{\sum_{i=1}^{n-1} (x_i - \bar{x})^2}, \quad (3.5)$$

where x_i are the data points and $\bar{x}$ is the mean of these values in the window X . The values of ρ for each time window measures the Pearson correlation between itself and the previous time window. Suppose the fluctuations in the original time series are truly random, we should expected to find low values of ρ . However, for a system approaching a critical transition, we should expect to find instances of increasing values ρ , as discussed in Chapter 1.2.2. We present our findings in Section 3.3.

3.2.2 Non-detrended Methods

One of the disadvantage when carrying out detrended fluctuation analysis is having to ensure that the time-series is stationary (detrended). To overcome this, we also briefly explored the idea of applying wavelets analysis [105, 106, 107] as comparison, since wavelet methods do not require detrending [108]. To get a basic intuition of how wavelet analysis works, we can imagine the convolutions between wavelets of different characteristic scales (frequencies) across a time-series. We choose wavelets of different scales, capturing the dynamics across the spectra of frequencies present in the time-series. Since the wavelet is localized at each time point, the wavelet analysis also gives us a good temporal localization of the power spectra.

For our comparison, we have chosen the Haar wavelet [109, 110], since its functional form is simple and known to be sensitive to sharp changes in the underlying time series [111]. Mathematically, the mother wavelet of the Haar wavelet can be written as,

$$\psi(t) = \begin{cases} 1 & 0 \leq t < \frac{1}{2}, \\ -1 & \frac{1}{2} \leq t < 1, \\ 0 & \text{otherwise,} \end{cases} \quad (3.6)$$

with a corresponding scaling function (father wavelet),

$$\varphi(t) = \begin{cases} 1 & 0 \leq t < 1, \\ 0 & \text{otherwise.} \end{cases} \quad (3.7)$$

Choosing the scales to be powers of 2, so that the set of Haar functions are,

$$\psi_{n,k}(t) = 2^{n/2} \psi(2^n t - k), \quad (3.8)$$

where $n, k \in \mathbb{Z}$ are the scaling and translation parameters respectively. We performed the continuous Haar transform on the housing price time series using MATLAB's built-in continuous wave transform, 'cwt' function, and to obtain its power spectrum. We discuss our findings in Section 3.3 below.

3.3 Results and Discussion

We start by presenting two indicators, $\Omega_{1/2}$ and lag-1 autocorrelation, ρ , which was used to observe the effects of critical slowing down the Singapore housing market from the 1995 to 2014. We begin by analyzing the effects of critical slowing down and spectral reddening in both the mean price and volume time-series of the Singapore Condominium housing market segment.

Figure 3.4 shows the time evolution of the half-max cumulative power $\Omega_{1/2}$ obtained using Equation (3.4). Since a concentration of power in the lower frequencies suggests that the time series is experiencing critical slowing down, we look out for time periods where $\Omega_{1/2} \leq 2$ for a consecutive 6 months period. In Figure 3.4, we see both the mean price and volume time-series exhibits a low $\Omega_{1/2}$ values for two time periods, with the first corresponding to the Asian Financial Crisis (during mid 1997) and the second corresponding to the 2007 Bubble in the Condominium segment of the Singapore housing market, which we have presented and discussed previously in Chapter 2.

Although both the volume and the mean price time series are potentially suitable for early warning studies, we find that $\Omega_{1/2}$ for mean price is more sensitive indicator compared to that for volume. Therefore, we encourage the use of the mean price to monitor and detect prolonged periods of small $\Omega_{1/2}$ and subsequently the use of the volume indicator as a confirmation signal.

In Figure 3.3 we show the lag-1 autocorrelation function ρ defined by (3.5) in Section 3.2. From the graph, we see the lag-1 autocorrelation function ρ for mean condominium prices in Singapore increasing from early 1996, preceding the period of Asian Financial Crisis, before peaking out in mid 1997. This coincided with the Asian Financial Crisis, which was sparked by the collapse of the Thai baht in July 1997. The value of ρ then continued to remain at elevated levels till early 1999. This result is in agreement to the spectral-reddening signals detected using the $\Omega_{1/2}$ indicator for the Singapore condominium market. Similarly, we also detect a rapid rise in the ρ indicator, approaching the 2007 condominium bubble. This is again followed by subsequent elevated levels of ρ between 2007 to 2009, also agreeing with

Figure 3.3: The Lag-1 Pearson ρ autocorrelation of the Singapore condominium housing segment for (a) Mean Price and (b) Volume. The condominium bubble shown in Chapter 2 in year 2007 can be seen as a period of increased autocorrelation (shown in black boxes). This result agrees with that shown in the $\Omega_{1/2}$ indicator.

the signals indicated by $\Omega_{1/2}$. This results demonstrate the viability of the $\Omega_{1/2}$ and ρ indicators as early warning indicators for the Singapore condominium market as it approaches a bubble phase.

In Figure 3.5, we also show the spectrograms of both mean price and volume and make comparisons to the two former detrended techniques. From the figure, we see both the 1997 Asian Financial Crisis and the 2007 Condominium bubble appearing as an impulse-like signal. Also, the results obtained from the continuous wavelet transform spectrogram provides a better temporal localization of bubble as compared to the $\Omega_{1/2}$ and ρ indicators. It is also interesting note that volume spectrogram provided a much stronger indication of the 2007 bubble as compared to the mean price spectrogram, which is in contrast to the $\Omega_{1/2}$ indicator which is more sensitive for mean price. As such, we have successfully shown here as a proof of concept, the viability of using non-detrended techniques such as the wavelet transforms in conjunction with detrended fluctuation analysis to identify bubbles.

We move on to present the $\Omega_{1/2}$ indicator for the mean price and volume time series of the HK residential housing market in Figure 3.6. From the figure, we see two time periods where both the mean price and volume indicators showed signs of

Figure 3.4: The half-max cumulative power, $\Omega_{1/2}$, of the Singapore condominium housing segment for (a) Mean Price and (b) Volume. The condominium bubble shown in Chapter 2 in year 2007 can be seen as a period of low $\Omega_{1/2}$ value (shown in black boxes).

critical slowing down. Similar to the observations made in Singapore condominiums, we observe the mean price time-series to be a more sensitive indicator, remaining at low levels for a longer extended period as compared to the volume indicator. Analyzing the time periods with small $\Omega_{1/2}$, we observed that the first time period corresponded to the sovereignty transfer of HK from the British to China, as well as the Asian financial crisis, which also impacted HK. We also see a second period starting from mid to end 2006 where the $\Omega_{1/2}$ in both mean price and volume indicating a possible bubble formation during the mid of 2006 in the HK residential housing market.

We also plot the lag 1 autocorrelation indicator ρ for both mean price and volume of the HK residential market in Figure 3.7. Interestingly, we do not observe significant increase and subsequent elevated autocorrelation levels in both mean price and volume approaching the 1997 sovereignty transfer, but instead see a relatively high ρ value from the starting period of the dataset in 1992, through to the sovereignty transfer in 1997. Since HK was established as a free market under the British prior to the transfer of sovereignty to China in 1997, the period drawing towards the sovereignty transfer casted a shadow of uncertainty on the long term continuity of

(a) Mean Price Spectrum

(b) Volume Spectrum

Figure 3.5: The spectrogram of Singapore condominium housing segment for (a) Mean Price and (b) Volume, obtained using the Haar Wavelet Transform (3.8). Both the Asian Financial Crisis and the condominium bubble of 2007 (shown in Chapter 2) can be evidently seen as region of increased power in both the Mean Price and Volume spectrogram.

existing housing policies under mainland China's rule stemmed from the difference in political ideologies. This uncertainty could be responsible for a high level of correlated noise in the HK housing market, which could explain why the rise of the ρ indicator associated to this transition was not evidently seen.

Nonetheless, we see the increase of the ρ indicator in both mean price (late 2005 onwards) and volume (mid 2006), corresponding to the second signal of the $\Omega_{1/2}$ indicator. By plotting the mean transaction price in each district in Figure 3.8, we see the prices increase in tandem during the period 2006 to 2010. Although this increase in price is not necessarily the formation of a bubble, this is highly possible if we consider the signals shown in the early warning indicators.

Another interesting point to note in the early warning indicators is the signal of the $\Omega_{1/2}$ indicator for mean price from the period starting early 2003 to mid 2004. We also see accompanying signals in the ρ indicators for both price and volume, both increasing from 2001 to 2003, before sustaining at high level through to mid 2004. This was however only seen in the $\Omega_{1/2}$ for volume for a very brief period (in the first few months of 2004). We look through the literature and find that this

Figure 3.6: The half-max cumulative power, $\Omega_{1/2}$, of the HK housing market for (a) Mean Price and (b) Volume. We see two periods of low $\Omega_{1/2}$ value for both indicators, corresponding to the 1997 Asian Financial Crisis and the 2006 HK housing bubble which are shown in black boxes.

corresponds to the 2004 Hong Kong residential property market bubble in [19].

3.4 Conclusions

In conclusion, we highlighted the potential of treating housing markets as complex system and examining them under the lens of complexity theory. We use the Singapore condominiums and HK housing market as a preliminary case study, but recognize that the methods presented in this study is generic and can be widely applied. Using these tools, we detect early warning signals relating to known bubbles in housing markets. We find these early warning indicators evident in both the mean price and volume time series, exhibited in the low values of $\Omega_{1/2}$, as well as the increase and subsequently sustained high levels of autocorrelation, ρ , for both the Singapore condominiums and HK housing markets.

A large part of these findings are still works in progress, and the next step is to apply early warning indicators, such as those that detect skewness and increase in variance of the fluctuations in housing markets to better confirm the findings. Furthermore, we would also be introducing statistical tests that can improve robustness

Figure 3.7: The Lag-1 Pearson ρ autocorrelation of the HK housing market for (a) Mean Price and (b) Volume. We see the two periods of high autocorrelation corresponding to the Asian Financial Crisis in 1997 and the 2006 HK housing bubble, which is also observed by the $\Omega_{1/2}$ indicator (shown in black boxes)

of the results. This can be done by creating null models through random resampling process, then applying the early warning indicators on both the null and empirical data, before testing the concordance [112] of each indicator on both sets.

Figure 3.8: The mean residential transaction price of homes in HK from the period 1992 to 2010. We see a correlated increase in price across all districts starting in 2006 and rising towards 2010.

Chapter 4

Finite Sample Corrections for Parameter Estimation and Significance Testing

4.1 Introduction

The current era of Big Data has ushered in a new way to look at Science — and that is letting the data speak for itself. Because of this, we are now much more concerned about empirical distributions than we have been in the past, and to check what the empirical distributions could be in statistically rigorous ways. In the past, statisticians commonly tested empirical data against the univariate normal distribution [113]. Some of these tests focus on the goodness-of-fit of higher order moments [114, 115, 116], while others compare the test statistics against an Empirical Distribution Function (EDF) [117, 118, 119, 120]. In 2011, Nornadiah and Yap performed a systematic comparison of Anderson-Darling (AD), Lilliefors, Kolmogorov-Smirnov (KS), and Shapiro-Wilk (SW) using numerical simulations and concluded that the SW test is the best, followed closely by the AD test for a given significance [121].

Among them, the KS and Lilliefors tests can also be applied to non-normal distributions. In fact, many real-world data do not follow normal distributions. For instance, many social systems are known to have power-law distributions [122]. These include the financial returns [29, 123, 124, 125], word count [126, 127], city

size [128, 129], home price [62, 60, 51], wealth and income [130, 131] distributions. One simple but naive way to detect a power law is to plot the data in log-log scale, fit it to a straight line and determine the goodness of fit. However, this simple method has three major flaws: (i) many distributions (eg. exponential, gamma, log-normal) can also look straight in a log-log plot, especially if the range of data is small; (ii) the goodness of fit only quantifies how well the fit is visually but does not tell us how plausible the fit is; and (iii) if our data looks straight in both log-log and semi-log plots, the goodness of fit values obtained from the two cannot be directly compared since they were obtained from plots of different scales. Clauset, Shalizi and Newman (CSN) address precisely these three points in their 2009 paper [69], and the test they proposed is now considered by many the gold standard in curve fitting. We shall describe the main idea of the CSN technique in greater mathematical detail in Section 4.3.

Since the CSN test can be applied across distributions, we also use it to fit data that appear exponentially distributed. On many occasions, we discovered that the exponential fits look good visually, but have significance values (p -value) much lower than fits of other data to power laws, even though the latter look visibly poorer (See Chapter 2). In fact, in the CSN paper where empirical data is tested against a power law (PL), log-normal, exponential (EXP), stretched exponential, and a power law with cut-off, the exponential distribution consistently performs poorer than the other distributions. This was also the case when Brzezinski tested the upper-tail wealth data for China, Russia, US, and the World using the CSN method [132]. In these papers, the data might truly be non-exponentially distributed, so it is not surprising the exponential fits fail. However, the low p -values for the visually convincing exponential fits to our data suggest that something fundamental was missed.

We start by introducing the real data sets that will be used in this Chapter in Section 4.2. Next, we describe the CSN test and illustrate the two issues associated with fitting data to distributions defined over $(0, \infty)$ in Section 4.3. The first of these is the *finite largest element effect* (FLE), due to the largest element in the data being finite. Second, we also encounter the *finite number of elements effect*

(FNE), due to the sample size dependence of the goodness-of-fit measures. These two *finite sample effects* are often neglected in tests of statistical significance. Using the insights gained from our theoretical treatment of the problem, we designed both the estimators and test statistic to account for the FLE and FNE effects in Section 4.4. Also, since real data is frequently polluted by noise, we discuss the impact of noise on the p -value, and propose a test statistic that accounts for noise in Section 4.5. Finally, in Section 4.6, we apply the adjusted test statistics on our real data sets and compare the p -values obtained against those from the CSN test. We have also appended in Appendix D.1 a list of variables used in this Chapter.

4.2 Data

4.2.1 Taiwan Housing Data

The Taiwanese housing price data used in this Chapter is identical to the data set used in Chapter 2 of this report. The data can be obtained from the Department of Land Administration (<http://plvr.land.moi.gov.tw/DownloadOpenData>). All housing transactions from 2010 are available. Following the work done previously in Chapter 2 we used only data from the Greater Taipei Area, containing Taipei city, New Taipei city, Keelung City and Taoyuan city, from 2010 to 2012 for the Zhu Zai Da Lous (which are condominium-type housing). Since the distribution contains statistically significant outliers at the tail, we choose to truncate the data set at 4.5×10^5 NT\$/ft², to avoid the scenario where the distribution is polluted by outliers.

4.2.2 Taiwan Income Data

The Taiwan income data set contains information on the survey of household income in Taiwan. The summary of its findings can be found at <http://win.dgbas.gov.tw/fies/index.asp> while the raw data set is available through a paid subscription which can be obtained by completing a form at <http://win.dgbas.gov.tw/fies/order.asp>. Following previous studies reviewed in [131], the income distribution

generally has an exponential body and power-law tail. We use the Clauset-Shalizi-Newman method for PL [69] to obtain the x_{min} of the PL part of the Taiwan Income distribution. Next, we truncate the full data set at this value $x_{min}^{(PL)}$, which leaves us with only the exponential part of the distribution.

4.2.3 SGX Normalized Returns

It is well known from previous works that the short interval (~ 5 mins) normalized returns of stock price indices follow a PL distribution [29, 124, 125]. We check this result on the the Straits Times Index (STI), which tracks the Singapore Stock Exchange (SGX), using the methods discussed in this Chapter. Since most of the public sources only provide historical data for the daily close price, we need to constructed the index using the tick-historical data in order to get fine-grained, short-interval (~ 5 mins) data. To construct the STI price time series, we choose the price for a particular interval by considering the tick price that is closest the end of the 5-min interval. We do this for all component stocks of the STI during the period 2009 to 2016. We note that our index price time series is not entirely similar to the published STI, since the STI is capital weighted. Since we do not have the capital information of the underlying company, we used a simple average weighting formula to construct the index. However, we do not expect this difference in weighting to change the PL characteristics of its return distribution, as many of the stock indices returns (cited above) is PL distributed but generally have different weighting formulas.

Next we obtain the set of returns $\mathbf{r} = r_1, r_2, \dots, r_N$, by calculating the returns in each 5-min interval using

$$r_t = \frac{p_t - p_{t-1}}{p_{t-1}}, \quad (4.1)$$

where r_t and p_t are the return and index price at interval t , and p_{t-1} is index price of the previous $(t - 1)$ interval. The normalized return $\mathbf{r}^{(n)}$ is then calculated using

$$r_t^{(n)} = \frac{r_t - \langle \mathbf{r} \rangle}{\sigma(\mathbf{r})}, \quad (4.2)$$

where $\langle \mathbf{r} \rangle$ and $\sigma(\mathbf{r})$ are respectively, the mean and standard deviation of $\mathbf{r}$.

4.3 Reexamining Significance Testing for Empirical Distributions

Sometimes we have reasons to believe that our large data sets may be described by well known distributions, such as the normal distribution, power law distribution, exponential distribution, and so on, but with best-fit parameter values that we need to determine. The most common methods used to perform this *parameter estimation* are the Maximum Likelihood Estimation (MLE) [133], Maximum Entropy Method (MEM) [134, 135, 136], Regressions [137], and direct or indirect computation of moments [138]. Since it is possible to fit any distribution to any data set, we need to compute its *goodness of fit*, which can be the KS distance [119], the coefficient of determination (R^2) and other forms of distance measure [139, 140].

In a recent statement, the American Statistical Association warned the scientific community that the p -value “was never intended to be a substitute for scientific reasoning” [141, para. 2], and outline six principles that can prevent its misuse [142]. A *Nature* commentary on this statement also added that “[r]esearchers should describe not only the data analyses that produced statistically significant results, . . . , but all statistical tests and choices made in calculations” [143, para. 3]. We heed the warning in this paper, but argue that when properly computed and interpreted, the p -value is useful in that it provides a quantitative and objective alternative to visual inspection of the fits. The latter is frequently subjective and biased. This utility becomes important when we are comparing fits of two or more data sets to two or more distributions, and have the ambiguity of being able to choose from two or more definitions of goodness of fit. This is why we need to go beyond the goodness of fits, to establish how plausible different distributions are for different data sets.

In 2009, Clauset, Shalizi and Newman (CSN) did precisely this by coming up with a p -test model that use the well-known PL distribution as an illustration. They started by writing down the probability density function for PL distribution

$$f_{PL} = \frac{\alpha - 1}{x_{min}^{1-\alpha}} x^{-\alpha}, \quad (4.3)$$

for $x \in [x_{min}, \infty)$. The CSN p -test involves four major steps:

4.3.1 CSN(i) MLE Estimation of α

Given an empirical data with S observations, $\mathbf{Y} = \{y_1, y_2, \dots, y_S\}$ sorted such that $y_i \leq y_{i+1}$, the CSN algorithm (**CSN(ia)**) first constructs the S subsets $\mathbf{X}^{(j)} = \{x_1^{(j)} = y_j, x_2^{(j)} = y_{j+1}, \dots, x_N^{(j)} = y_S\}$. (**CSN(ib)**) For each $\mathbf{X}^{(j)}$, we estimate $\alpha^{(j)}$ using the MLE method that maximizes the log-likelihood function,

$$\ln \mathbb{L}_{PL} = \ln \left[\prod_{i=1}^N f_{PL}(x_i | \hat{\alpha}) \right] = N \ln \left(\frac{\hat{\alpha} - 1}{x_{min}} \right) - \hat{\alpha} \sum_{i=1}^N \frac{x_i}{x_{min}}. \quad (4.4)$$

Applying the maximizing condition $\frac{\partial(\ln \mathbb{L})}{\partial \alpha} = 0$ yields,

$$\hat{\alpha} = 1 + \left\langle \ln \frac{x}{x_{min}} \right\rangle^{-1}, \quad (4.5)$$

where the hat indicates an estimated parameter. We can also check that the second derivative is,

$$\frac{\partial^2[\ln \mathbb{L}]}{\partial \alpha^2} = -\frac{1}{(\hat{\alpha} - 1)^2} < 0, \quad (4.6)$$

which is the condition for $\mathbb{L}$ to be maximized.

4.3.2 CSN(ii) KS Distance

If the set $\mathbf{X}$ follows probability distribution function f_X with CDF F_X , then its probability integral transform (See Appendix C.1) $u = F_X(x)$ is a standard uniform distribution function ($U(0, 1)$). For the PL distribution this is,

$$F_{PL}(x) = 1 - \left(\frac{x}{x_{min}} \right)^{1-\alpha}. \quad (4.7)$$

where $F_{PL}(x)$ is the CDF of the power law distribution and with inverse CDF,

$$F_{PL}^{-1}(u) = x_{min}(1 - u)^{\frac{1}{1-\alpha}}. \quad (4.8)$$

For any PL distributed sample $\mathbf{X} = \{x_1 = x_{min}, x_2, \dots, x_N\}$ with estimated $\hat{\alpha}$, we (**CSN(ia)**) first transform the sample to $U^{(s)} = \{u_i^{(s)} = F_{PL}(x_i | \hat{\alpha})\}_{i=1}^N$ using (4.7). (**CSN(iib)**) Then we calculate the KS distance

$$d_{KS} = \forall_{i=1}^N \sup \left(\left| u_i - \frac{i}{N} \right| \right) \quad (4.9)$$

between $U^{(s)}$ and $U(0, 1)$. Here we make use of the fact that the CDF of $U(0, 1)$ is a linear function, $F_U(u) = u$.

4.3.3 CSN(iii) Determining x_{min}

To determine x_{min} , (CSN(iii)) we calculate the KS distance for each subset $\mathbf{X}^{(j)}$ with its corresponding $\hat{\alpha}^{(j)}$. (CSN(iii)) The set $\mathbf{X}^{(j)}$ that yields the lowest KS distance ($d_{KS}^{(em)}$) gives us $\hat{x}_{min}^{(em)} = y_j$ and $\hat{\alpha}^{(em)} = \hat{\alpha}^{(j)}$. The superscript ‘(em)’ indicates a parameter obtained from empirical data.

4.3.4 CSN(iv) Significance Testing

After $\hat{\alpha}^{(em)}$ and $\hat{x}_{min}^{(em)}$ have been estimated from $\mathbf{Y} = \{y_1, y_2, \dots, y_S\}$, we test how plausible it is for $\mathbf{X} = \{x_1 = \hat{x}_{min}^{(em)}, x_2, \dots, x_N\} \subset \mathbf{Y}$ to be a sample taken from a PL distribution. This is done by (CSN(iva)) sampling the PL M times using $\hat{\alpha}^{(em)}$ and $\hat{x}_{min}^{(em)}$. (CSN(ivb)) For the m^{th} simulated sample we go through CSN(i) to CSN(iii) to obtain $d_{KS}^{(m)}$. (CSN(ivc)) The significance measure

$$p = \frac{1}{M} \sum_{m=1}^M \mathbb{I}_{\{d_{KS}^{(em)} < d_{KS}^{(m)}\}}. \quad (4.10)$$

is the fraction of simulated samples whose fits are poorer than that of the data.

Extending the CSN method to other distributions, we performed p -testing on the Taiwan home price per square foot (fitted to EXP), Taiwan income (fitted to EXP), and the Straits Times Index normalized return (fitted to PL). The fits and p -values are shown in Fig. 4.1. All fits are visually good yet only the p -value for the Taiwan housing is appreciable. We realized the reason for this is simple: while both the EXP and PL distributions are defined over $(0, \infty)$, when we collect data from the real world we can only obtain a finite number of elements. Moreover, the largest element in the data is finite. However, existing tests for statistical significance generally do not account for the effects produced by having a finite number of elements (FNE) and a finite largest element (FLE). In the next section we will explain how the parameters and test statistics can be adjusted for FNE and FLE.

Figure 4.1: p -test and fits of (a) 2012 to 2014 Taiwan home price per square foot in order of $\times 10^5$ (fitted to EXP), (b) 2012 Taiwan lower-tail income in million (fitted to EXP), and (c) 2009 to 2016 Straits Time Index normalized return (fitted to PL). The black dots represent empirical data while the blue dashed line represents the fit. All fits are visually good, yet only the p -value (P_{KS} in percentage) for Taiwan home price is appreciable.

4.4 Finite-Sample Adjustments

4.4.1 Parameter Adjustment for Finite Largest Element

Here, we will illustrate the effects of FLE using an asymptotic EXP distribution. Apply the same methodology, the same discussion can be generalized to other distributions. We have also included the detailed derivations for both the EXP and PL distributions in Appendix D.2.

The EXP distribution is defined as

$$f_{EXP}(x) = \beta \exp[-\beta(x - x_{min})], \quad (4.11)$$

with β as a sole parameter for $x \in [x_{min}, \infty)$. Maximizing the likelihood function $\mathbb{L} = \prod_{i=1}^N P(X = x_i | x_{min}, \hat{\beta})$, we find the estimated parameter

$$\hat{\beta} = \frac{1}{\langle x \rangle - x_{min}}. \quad (4.12)$$

We also check that the second derivative is,

$$\frac{\partial^2 [\ln \mathbb{L}]}{\partial \beta^2} = -\frac{1}{\hat{\beta}^2} < 0 \quad (4.13)$$

which is the condition for $\mathbb{L}$ to be maximized.

If we use the mean obtained from data $\langle x \rangle_{data}$ as $\langle x \rangle$ in (4.12) we will obtain the unadjusted estimator β_{unadj} . However due to the FLE, we can only average up till x_{max} . As such $\langle x \rangle_{data}$ will be biased downwards and (4.12) over-estimates $\hat{\beta}$.

To adjust for the FLE, we add the truncated part back into $\langle x \rangle_{data}$, to define the adjusted $\langle x \rangle_{adj}$ as

$$\begin{aligned} \langle x \rangle_{adj} &= \langle x \rangle_{data} \int_{x_{min}}^{x_{max}} f_{EXP}(x) dx + \int_{x_{max}}^{\infty} x f_{EXP}(x) dx \\ &= \langle x \rangle_{data} \{1 - \exp[-\beta(x_{max} - x_{min})]\} + \frac{\exp[-\beta(x_{max} - x_{min})]}{\beta} [\beta x_{max} + 1] \end{aligned} \quad (4.14)$$

Inserting $\langle x \rangle_{adj}$ for $\langle x \rangle$ into (4.12), we obtain a nonlinear equation

$$\left[\hat{\beta}_{adj} (x_{max} - \langle x \rangle_{data}) + 1 \right] \exp \left[-\hat{\beta}_{adj} (x_{max} - x_{min}) \right] + \hat{\beta}_{adj} (\langle x \rangle_{data} - x_{min}) - 1 = 0 \quad (4.15)$$

that we need to solve to obtain $\hat{\beta}_{adj}$. This inclusion of x_{max} is an extension of the series of work [144, 145, 146] on improving parameter estimation for PLs given both x_{min} and x_{max} .

To test the performance of this adjustment formula, we simulated 1000 sets of EXP distributed data for $10^{-4} \leq \beta_T \leq 10^2$, by using the inverse cumulative function for EXP distribution

$$F_{EXP}^{-1}(u, \beta_T) = x_{min} - \frac{1}{\beta_T} \ln(1 - u). \quad (4.16)$$

This transforms $U(0, 1)$ distributed elements $\{u_i\}$ to EXP distributed elements $\{x_i\}$ (see Appendix C.1). Using this transformation F_{EXP}^{-1} , 0 and 1 map to x_{min} and ∞ respectively. It is also useful to note that (4.16) is the inverse of the CDF of the EXP distribution,

$$F_{EXP}(x, \beta_T) = 1 - \exp[\beta(x_{min} - x)]. \quad (4.17)$$

To simulate the effect of a FLE with $x_{max} = F_{EXP}^{-1}(0.9)$, we sampled 1000 sets of EXP distributed data using $U(0, 0.9)$ instead of $U(0, 1)$ with $x_{min} = 0$. Thereafter, we estimated $\hat{\beta}_{unadj}$ and $\hat{\beta}_{adj}$ using (4.12) and (4.15). Fig. 4.2 shows the relative

estimation errors

$$\Delta\beta = \frac{\sqrt{\langle (\hat{\beta} - \beta_T)^2 \rangle}}{\beta_T} \quad (4.18)$$

of $\hat{\beta}_{unadj}$ and $\hat{\beta}_{adj}$ with respect to the true beta β_T . As we can see from the Fig. 4.2, $\Delta\hat{\beta}_{unadj}$ is about 38% for small samples $N \sim 10^2$ and decreases to 34% for large samples $N \sim 10^4$. On the other hand, $\Delta\hat{\beta}_{adj}$ starts at 20%, but decreases to 2% as the number of data points is increased. Although it can be shown that the $\hat{\beta}_{unadj}$ is an unbiased estimator [147, 69], we find it converging very slowly with increasing sample size in the unfortunate situation of a small x_{max} . In contrast, $\hat{\beta}_{adj}$ converges very quickly even for small x_{max} as we have accounted for the FLE.

Figure 4.2: Relative estimation errors of (a) $\hat{\beta}_{unadj}$ and (b) $\hat{\beta}_{adj}$ measured from 1000 simulated samples using different β_T and N with $x_{min} = 0$ and $x_{max} = F_{EXP}^{-1}(0.9)$. Due to the FLE, $\Delta\beta_{unadj}$ remains high (close to the theoretical relative error of $\epsilon(\delta = 0.1, x_{min} = 0) = 0.1 [1 - \ln(0.1)] \approx 33\%$) even for large N . In contrast, $\Delta\beta_{adj}$ decreases rapidly with increasing N .

In Appendix D.3, we show details for our derivation of the theoretical relative error

$$\begin{aligned} \hat{\beta} &\approx \beta_T + \beta_T [\beta_T x_{max} + 1] \exp(-\beta_T(x_{max} - x_{min})) + \\ &\quad \mathcal{O}\{(\beta_T x_{max} + 1)^2 \exp(-2\beta_T(x_{max} - x_{min}))\}. \end{aligned} \quad (4.19)$$

In (4.19), we see that the relative error is not explicitly related to the sample size. Also, we can conclude that the $\hat{\beta}_{unadj}$ is always larger than the β_T because of

the FLE effect. The convergence rate then depends on how rapidly x_{max} approaches infinity with increasing sample size.

4.4.2 Test Statistic Adjustment for FLE

For a finite sample, $F_{EXP}(x) < 1$ for all $x < \infty$. Mathematically, this means that $F_{EXP}(x) \sim U(0, 1 - \delta)$, where $F_{EXP}^{-1}(x_{max}) = 1 - \delta$. This observation is important, because d_{KS} is obtained by comparing $U^{(s)} = \{u_i^{(s)} = F_{EXP}(x_i|\hat{\beta})\}_{i=1}^N$ against $U(0, 1)$ (see (4.9)). This tell us that for a fair comparison, we need to rescale all elements in $U^{(s)}$ by a factor of $1/(1 - \delta)$. Fig. 4.3 shows the d_{KS} measured for the 1000 sets of EXP distributed data with finite largest element $x_{max} = F^{-1}(0.9)$ for various β_T and sample sizes N . For each sample, we use (4.15) to estimate the $\hat{\beta}_{adj}$ and transformed this data to $U^{(s)}$ using (4.17). After that, we measure d_{KS} with (4.9) to obtain unadjusted KS distance, KS_{unadj} and adjusted KS distance, KS_{adj} using the non-rescaled and rescaled $U^{(s)}$ respectively. KS_{unadj} goes from 0.14 for small samples $N \sim 10^2$, to 0.10 for large samples $N \sim 10^5$. In contrast, KS_{adj} decrease from 0.06 for small samples to 0.006 for large samples.

Figure 4.3: The median KS distances for (a) KS_{unadj} and (b) KS_{adj} measured from 1000 simulated samples using different β_T and N . The x_{min} is set to 0 and $x_{max} = F_{EXP}^{-1}(0.9)$. Because of the FLE, KS_{unadj} remains above $\delta = 0.10$ while KS_{adj} converges to zero for large N .

4.4.3 Adjustment for Finite Number of Elements

Until now, we have only discussed adjustments to the estimated parameter and the KS distance to eliminate the bias caused by the FLE. Besides the FLE effect, we also need to consider the bias caused by having a finite number of elements in the sample. As we can see from Fig. 4.3, the KS distance decreases as the sample size increases. Therefore, in order to have a fair comparison of the goodness of fit for various sample sizes, we need to determine how d_{KS} changes as a function of N . To do this, we simulated 10^6 samples of various sizes N from $U(0, 1)$. For each sample we determined d_{KS} using (4.9), so that for each N we end up with 10^6 KS distances. In Fig. 4.4 we show the KS distances at different deciles, which exhibits the asymptotic behavior

$$d_{KS}(\varphi_{KS}, N) = \frac{\left(\frac{100}{\varphi_{KS}} - 1\right)^{-0.176} \exp(-0.274)}{N^{0.492}}, \quad N > 50 \quad (4.20)$$

that we settled for, after experimenting with several functional forms (see Appendix D.4). This result agrees with our expectation that $d_{KS} \rightarrow 0$ as $N \rightarrow \infty$. It also suggests that if we have two samples with sizes N_1 and N_2 from the same distribution, we should compare $N_1^{0.492} d_{KS}^{(1)}$ against $N_2^{0.492} d_{KS}^{(2)}$. Otherwise, if $N_2 > N_1$ then naturally $d_{KS}^{(2)} < d_{KS}^{(1)}$ and we will be lead to the wrong conclusion that the N_2 sample fits the distribution better.

4.5 The Effects of Random Noise

Besides having to work with finite samples and finite largest elements, we will also in practice encounter imperfections while collecting samples for various reasons, such as undetected samples, contamination by background noise, and recording errors. We call such noises that occur at the element level *elementary noise*. When we convert these samples to a distribution, noise will also be present at the distribution level that we refer to as *distribution noise*. In principle the information at the distribution level is more robust compared to the elementary level, as we expect random and thus uncorrelated noise to cancel each other. This means that the distribution is less sensitive to elementary noise, but we still worry whether the distribution noise

Figure 4.4: Log-log plot of d_{KS} against N for different deciles going from the 10th percentile (blue) to the 90th (red), obtained from 10⁶ simulations.

may play an important role in our test of statistical significance. In order to account for the effects of distribution noise, we need to first be able to quantify distribution noise, and thereafter understand how it affects significance testing.

Suppose we now randomly generate a set of EXP data. After adjusting for FLE, we obtained the distribution parameters and use it to transform this set to $U^{(s)} = \{u_i^{(s)} = F_{EXP}(x_i|\hat{\beta})\}_{i=1}^N$ following the procedure outlined in Section 4.4.1. Then as illustrated in Fig. 4.5(a)-(b), a natural way to measure the distribution noise is to plot the histogram, count the frequency for each bin, and compare it to the expected frequency from $U(0, 1)$. Since this can be more accurately done for smaller bin sizes, we use the intervals between sorted elements as a collection of non-uniform bins, as shown in Fig. 4.5(c)-(d). For a data set consisting of N elements, each bin carry a weight of $1/N$, evenly distributed within the interval $(u_{i-1}, u_i]$, such that the probability density is

$$f(u_{i-1}, u_i) = \frac{\frac{1}{N}}{u_i - u_{i-1}}. \quad (4.21)$$

As the theoretical probability density for $U(0, 1)$ is 1, we define the distribution

noise d_{DN} mathematically to be

$$d_{DN} = \sqrt{\frac{\sum_{i=1}^N (u_i - u_{i-1})^2 [f(u_{i-1}, u_i) - 1]^2}{\sum_{i=1}^N (u_i - u_{i-1})^2}} = \sqrt{\frac{\sum_{i=1}^N (u_i - u_{i-1})^2 \left(\frac{1}{N(u_i - u_{i-1})} - 1 \right)^2}{\sum_{i=1}^N (u_i - u_{i-1})^2}}, \quad (4.22)$$

where $u_0 = 0$ and $u_N = 1$. We need to weigh the deviation of each bin by $(u_i - u_{i-1})^2$ because the bins are non-uniform. and to keep d_{DN} finite.

Figure 4.5: Illustration of the distribution noise we would measure if we sample 10 elements from $U(0,1)$, rescaled such that the largest element becomes 1. In (a)-(b) we use 5 uniform bins whereas in (c)-(d) we use the intervals between sorted elements as the bins. The counts are shown on the left, whereas the probability densities calculated using (4.21) are shown on the right.

4.5.1 Relation between Distribution Noise and Sample Size

As with Section 4.4.3, we simulated 10^6 samples from $U(0,1)$ with different N . For each sample, we calculate the distribution noise d_{DN} using (D.13) and plot its deciles against N as shown in Fig. 4.6. After experimenting with several functional forms, we write down the relationship between d_{DN} and N at percentile $\wp_{DN}$ as

$$d_{DN}(\wp_{DN}, N) = \langle d_{DN} \rangle + \Phi(\wp_{DN} - 50) \frac{\exp\left(-\frac{[50 - |\wp_{DN} - 50|]^{0.430}}{|\wp_{DN} - 50|^{0.302}}\right)}{N^{0.495}}, \quad (4.23)$$

where $\Phi(x)$ represents the sign of x , and

$$\langle d_{DN} \rangle = \sqrt{\frac{1}{2} + \frac{2-N}{2N^2}} \left(\frac{N}{N+0.5} \right) \quad (4.24)$$

is the analytically derived distribution noise, that converges to $1/\sqrt{2}$ as $N \rightarrow \infty$ (refer to Appendix D.4 for more details). This result suggests that if we have two samples with sizes N_1 and N_2 with $N_2 > N_1$ from the same distribution, we should compare $N_1^{0.495}(d_{DN}^{(1)} - 1/\sqrt{2})$ against $N_2^{0.495}(d_{DN}^{(2)} - 1/\sqrt{2})$. Otherwise, we risk making wrong conclusion that the N_2 sample fits the distribution better if $d_{DN}^{(1)} > d_{DN}^{(2)}$.

Figure 4.6: Relationship between distribution noise d_{DN} and sample size N at deciles going from the 10th percentile (blue) to the 90th (red), obtained from 10^6 simulations. The d_{DN} value converges to $1/\sqrt{2}$ as N increases.

4.5.2 Relationship between Distribution Noise and KS Distance

As measures for statistical deviations, d_{DN} and d_{KS} are different in that d_{DN} measures deviation at the probability density level, whereas the d_{KS} measure it at the cumulative density level. As a result, d_{KS} assigns more weight to the tail of the distribution, while d_{DN} is more sensitive to deviations in the body of the distribution.

Therefore, if we wish to combine these two measures to estimate the significance level, we need to first investigate the relationship between d_{KS} and d_{DN} . We do this by simulating 10^6 samples from $U(0, 1)$ for various sample sizes, and for each sample, we calculate d_{KS} and d_{DN} using (4.9) and (D.13) respectively, to obtain 10^6 pairs of d_{KS} and d_{DN} . We then compute the Pearson correlation between d_{KS} and d_{DN} and learned that (see Appendix D.4 for the comparison of fits)

$$\rho_{d_{KS}, d_{DN}}(N) = \frac{e}{N^{0.481}}. \quad (4.25)$$

As expected, d_{KS} is positively correlated with d_{DN} . Since d_{KS} is a measure at the cumulative level, the random distribution noises cancel each other, thus the correlation between d_{KS} and d_{DN} vanishes as $N \rightarrow \infty$.

4.6 Application to Significance Testing

4.6.1 Significance Level for a Given Distribution

To perform significance testing given d_{KS} and d_{DN} , we need the percentile values $\wp_{KS}$ and $\wp_{DN}$. $\wp_{KS}$ can be obtained by inverting (4.20), as

$$\wp_{KS}(d_{KS}, N) = \frac{100}{\left(1 + (d_{KS} N^{0.492} \exp(0.274))^{-\frac{1}{0.176}}\right)}. \quad (4.26)$$

Similarly, we invert (4.23), and solve

$$\begin{aligned} \wp_{DN}^{0.430} + (50 - \wp_{DN})^{0.302} \ln(|\eta| N^{0.495}) &= 0, \quad \eta < 0 \\ \wp_{DN} - 50 &= 0, \quad \eta = 0 \\ (100 - \wp_{DN})^{0.430} + (\wp_{DN} - 50)^{0.302} \ln(|\eta| N^{0.495}) &= 0, \quad \eta > 0 \end{aligned} \quad (4.27)$$

to get $\wp_{DN}$, where $\eta = d_{DN} - \langle d_{DN} \rangle$.

Substituting the empirical KS distance $d_{KS}^{(em)}$ and empirical distribution noise $d_{DN}^{(em)}$ into (4.26) and (4.27), we obtain $\wp_{KS}^{(em)}$ and $\wp_{DN}^{(em)}$. This is an alternative way of obtaining the p -value without the need to perform Monte-Carlo (re)sampling again (CSN method), since we have already done so in Section 4.4 and 4.5. The percentage of simulated $U(0, 1)$ samples with $d_{KS/DN} > d_{KS/DN}^{(em)}$ is $100 - \wp_{KS/DN}^{(em)}$. Since d_{KS}

and d_{DN} are not independent ((D.22)), we discount the correlation between d_{KS} and d_{DN} , and define the significance level (p -value) as

$$p(\wp_{KS}, \wp_{DN}, N) = \sqrt{\left(1 - \frac{\wp_{KS}}{100}\right) \left(1 - \frac{\wp_{DN}}{100}\right) \left(1 - \frac{e}{N^{0.481}}\right)}, \quad (4.28)$$

to avoid overestimating the significance level.

4.6.2 Fitting to Empirical Data

We follow the steps outlined in the CSN algorithm (Section 4.3) to fit the empirical data, but with two important modifications: (Ii) the parameters (**CSN(ib)**) and goodness of fit (**CSN(iib)**) are adjusted for the finite largest element; and (Iii) the p -value (**CSN(ivc)**) is adjusted for the finite number of elements effect. Meanwhile, optional modifications are (Oi) to incorporate distribution noise as another dimension for goodness of fit, so that the p -value can be determined via $d_{KS}^{(em)}$, $d_{DN}^{(em)}$, or both; (Oii) instead of using bootstrapping to determine the p -value in the CSN method, which is very slow for large samples, one can use the fast inversion formulae (4.26), (4.27), or (4.28).

Fig. 4.7 shows the fits and p -testing results for Taiwan housing price, Taiwan wealth, and Straits Times Index normalized returns. It is reassuring that after modifications the p -values of all distributions increased. In particular, the two distributions (Fig. 4.7(b),(c)) that did not meet the $p > 0.1$ criterion (as suggested by [69]) before modification, now have $p > 0.5$. This is in agreement with our visual assessment of the three fits. We also understand now that a large δ (small x_{max}) is the main reason for Taiwan wealth to fail p -testing before adjustment (although the fit is visually good). In general, our correction formulas perform the best when δ is large due to small sample sizes or truncations. In Appendix D.5 we show more plots and instances where small δ values affects the significance testing.

There are several limitations one should note while obtaining $P_{KS/DN}$ using (4.26) or (4.27). First, it is only applicable to large samples (see Figs. 4.4 and 4.6). Second, these equations are obtained after experimenting with several functional forms and are only approximate. Lastly, p_{KS} measured using the CSN method are consistently smaller than that based on (4.26). This is due to the CSN algo-

Figure 4.7: p -testing results for (a) 2012 to 2014 Taiwan home price per square foot in order of $\times 10^5$ (fitted to EXP), (b) 2012 Taiwan lower-tail income in millions (fitted to EXP), and (c) 2009 to 2016 Straits Times Index normalized return (fitted to PL) before and after finite-sample adjustments. In this figure, the empirical CDF that is adjusted for FLE is shown as black dots, while the unadjusted and adjusted fits are shown as blue and red dashed line respectively. $P_{KS/DN}$ -values (in percentage) are for unadjusted (blue) and adjusted (red) fits. We separate the p -values obtained using the CSN method (left) from those using (4.26) or (4.27) (right) by a '/'.

rithm having an extra step to select x_{min} that minimizes d_{KS} of each simulated sample, and thus the algorithm is stricter than our fast inversion formulae. However the inversion formulae (4.26) and (4.27) is convenient and provides an upper bound for $P_{KS/DN}$. We make the codes for these two methods available at <https://github.com/BoonKinTeh/StatisticalSignificanceTesting> but leave it to the reader to decide which method to use.

All in all, when we test for statistical significance, we need to be aware of finite sample effects, namely the finite largest element effect and the finite number of elements effect. Beyond the KS distance measured at the cumulative distribution level, we also introduce an alternative measure of the goodness of fit based on the distribution noise at the probability density level.

4.7 Conclusion

In conclusion, we addressed one of the increasingly pressing issues in Big Data, which is fitting empirical data sets to distributions. In this Chapter, we extend the work of Clauset *et al.* to note that all real data sets are finite, and hence they suffer from finite sample effects. Such effects are especially pronounced in truncated data sets. Using real data sets as examples, we show that these effects cannot be neglected, as they can lead us to sometimes understate the statistical significance of the fits and reject fits that are visibly good. Hence, we outline a procedure for curve fitting with modifications that accounts for these effects, allowing us to compare fits across different distributions and different sample sizes. Indeed, after the changes we see significant improvement in the statistical significance of the visually good fits, reducing the risk of mistakenly rejecting plausible fits. This will benefit all disciplines fitting large data sets to several distribution functions, and use the best candidate to guide the development of theories and models.

Chapter 5

Conclusions

5.1 Summary

To summarize the works presented in this thesis, we showed in Chapter 1 the motivation and importance of studying housing markets. We reviewed past works and found two research gaps, which are addressed in this thesis. The first being the lack of clear and commonly agreed upon definition of housing bubbles. The second gap pertains to lack of generic methods in the study of bubbles in housing market, which are complex systems, so factors that causing housing bubbles cannot be studied in isolation.

In Chapter 2, we identified that the equilibrium distribution in housing markets is Gibbs-Pareto for the housing markets we study. Next, we show that housing bubbles are positive departures from the equilibrium distribution and manifests themselves as DKs. Using the U-test developed by Pisarenko and Sornette [70] to test for these DKs, we are able to attach statistical significance to these positive outliers. This gives us an unambiguous statistical definition of housing bubbles. Using this, we show the presence of a bubble in the condominium segment of the Singapore housing market. Also, we identified the formation of a bubble in the Zhu Zai Da Lous in the Greater Taipei Area during the period 2010 to 2012. Further analyzing the spatial and temporal dynamics of the bubble in the Singapore condominium market, we see that the bubble began to form in mid 2007, starting in the prestigious postal districts of 9 and 10 before diffusing spatially to the peripheral districts.

Following on the findings in the Singapore housing market, performed time series analysis in Chapter 3 to identify generic early warning indicators associated with critical slowing down and spectral reddening in the Singapore housing market. Doing so, we find that early warning signals associated with both the 2007 condominium bubble, as well as the 1997 Asian Financial Crisis can be detected. Applying similar indicators to the Hong Kong residential housing market, we find early warning indicators corresponding to a bubble from 2006 to 2010.

A large part of our work in this thesis requires us to fit data to empirical distributions. Since these fits determines the equilibrium distributions of the housing data, it is important that the parameter estimation and the feasibility of the fits are statistically significant. In many cases, when we fit data to exponential distributions, we often see fits that are visually good, but have low or no significance. Reviewing the literature in curve fitting and parameter estimation, we realized that the low significance in fitting is a result of finite sample effects in real data. To account for this, we derived correction formulas for these finite size effects and showed both theoretically and computationally that this greatly improves the parameter estimation and statistical significance for noisy real world data. This led in turn to improved statistical significance of the estimated parameters for equilibrium housing distribution obtained from data.

5.2 Future Works

5.2.1 Housing Locale Theory

The works presented in this thesis have provided us with an understanding of empirical distributions in housing markets. A natural step moving forward would be to understand the mechanisms in housing markets, which could lead to these distributions. One toy model that is worth exploring is the housing locale theory, in which a city center acts as an attractor for people to congregate. If an agent live at a home situated at the city center, he derives a maximum amount of utility, U_0 . Alternatively, he can choose to stay a distance r away, pays a cheaper rent $P(r) = P_0 r^{-\alpha}$ to stay away from the city, but pays a time penalty and traveling cost $R(r) = R_0 r^\beta$

to visit the city center. Hence, the utility of such an agent is,

$$U(r) = U_0 - \frac{P_0}{r^\alpha} - R_0 r^\beta. \quad (5.1)$$

The at equilibrium, we have the condition,

$$\frac{dU}{dr} = \frac{\alpha P_0}{r^{\alpha+1}} - \beta R_0 r^{\beta+1} = 0. \quad (5.2)$$

However, this yields a simple equilibrium solution, $r_* = \left(\frac{\alpha P_0}{\beta R_0}\right)^{\frac{1}{\alpha+\beta}}$. In real housing markets, we cannot have the case where every single agent are stuffed within a single ring r_* away from the city, since this depends on the number of homes, N within that ring. Assuming the density of homes σ , is constant so the number of homes is directly proportion to the area, we should find more homes per unit distance away from the city center, since $dN = 2\sigma\pi r dr$. Therefore, even if an agent wants to sell its home and move closer to the city center, it must also consider whether there are enough homes closer to the city center to accommodate this shift.

Thermodynamically, this increasing in number density is related to concept of entropy,

$$S(r) = k \log \Omega(r), \quad (5.3)$$

where k play the role of the Boltzmann constant, and $\Omega(r)$ must be computed from $N(r)$ using combinatorics. The larger the entropy, the lesser the available utility, and the less expensive homes should be. Therefore, we need to consider the free energy,

$$U(r) = U_0 - \frac{P_0}{r^\alpha} - R_0 r^\beta - TS(r). \quad (5.4)$$

where T plays the role of temperature in the housing market. Here, the home price $P(r)$ can be determined by the similar equilibrium condition, $\frac{dU}{dr} = 0$. However, given the additional entropy term, with its relation to $N(r)$, we expect an exponential distribution to emerge in the home price distribution.

5.2.2 Housing Portfolio Theory

The two part Gibbs-Pareto distribution found in both income/wealth distributions and home price distributions suggest the close relationship between wealth and home

prices. Here, we want to understand how agents (families) make housing transactions to optimize their housing portfolios.

The families from the exponential part of the wealth distribution can usually only afford one home. If all families use the same proportions of their wealth to buy houses, the rich from the Pareto part of the wealth distribution will either (i) own a more expensive house or (ii) own more houses. To cater to case (i), this would mean that we need to have big houses in expensive areas, which is generally not seen in space-constrained cities. Furthermore, homes fulfill the role of bringing the members of the families together, and too big a house would be contradicting to this. This leaves point (ii), so that agents create portfolios of houses instead. Each agent can own a portfolio of houses, $\vec{P} = [P_1, P_2, \dots, P_n]$ so that the total wealth, $W = \sum_{i=1}^n P_i + w$ is the combined value of these homes and a disposable wealth w . Here, we assume that the only instrument that the agents can invest in are houses in the housing market.

The agents can then sell the i -th home at $(1 + \rho_i)P_i$ for a return of ρ_i . This increases its disposable wealth $w \rightarrow w + (1 + \rho_i)P_i$.

Agents can also purchase homes if they have enough disposable wealth to do so, such that,

$$w \rightarrow w - p_{n+1}, \quad \vec{P} = [P_1, P_2, \dots, P_n, P_{n+1}],$$

Since disposable wealth generates no return, agents should try to use up w until such time when all agents have $w = 0$. In this equilibrium theory, we cannot talk about the expected return of the i -th home as if it is a stock, because the housing market is an illiquid market. Suppose we start off with no location dependence, then all homes are equivalent, so each P_i follows the same distribution. If we further assume that the agent decides very early on what it wants $\rho_i = \rho$ to be, then the question is not about at what price, but at what time the transaction will occur. Logically, a smaller ρ tends to decrease the average time between transaction. More importantly, an agent with more homes in its portfolio will have a smaller time, τ , between transactions, for the same return ρ .

Assuming that the price of each home fluctuates and performs a random walk, then the average rate of return can be described as a *first passage problem*. For

such Brownian particle, this is just the Levy distribution, which is the result first shown by Schrödinger and Smoluchowski [148]. We can then calculate the mean and variance for the rate of return $\frac{\rho}{\tau}$,

$$\langle \dot{\rho} \rangle = \int_0^\infty \frac{\rho}{\tau} f_L(\tau) d\tau, \quad (5.5)$$

$$Var(\dot{\rho}) = \int_0^\infty \left(\frac{\rho}{\tau}\right)^2 f_L(\tau) d\tau, \quad (5.6)$$

where $f_L(\tau)$ is the Levy distribution.

However, the variance of the rate of return is not the true measure of risk, because the probability of a desired return ρ not being realized, is not negligible even for small values of $\rho > 0$. Therefore, we can talk about optimizing investing in a single home by maximizing the expected rate of return against the risk of *not* realizing the return, for portfolios with different number of homes n .

Furthermore, we can bring location dependence back into the housing portfolio theory, by adding a second random attribute, distance r to the city centre, to the homes in the housing portfolio. In the housing locale theory, we have assumed a dependence $P_0 r^{-\alpha}$ for home prices. Using the equilibrium of the spatial housing portfolio theory, we hope to derive a self-consistency condition that would allow us to fix the exponent α . For different choices of α , we can obtain different optimum housing portfolios with different home price distributions and dependences on the distance r away from the city centre. Using this, we hope to be able also explain the tail of the home price distributions we observe in Chapter 2.

5.2.3 Agent Based Simulations

The proposed theoretical works can also be simulated by building agent based model that take into account the mechanics for both the housing locale theory, as well as the portfolio theory. This will allow us to simulate agent interactions, which will lead to a much greater understanding of the microscopic dynamics of housing markets.

Bibliography

- [1] Kenneth A Froot, David S Scharfstein, and Jeremy C Stein. Herd on the street: Informational inefficiencies in a market with short-term speculation. *The Journal of Finance*, 47(4):1461–1484, 1992.
- [2] Rama Cont and Jean-Philippe Bouchaud. Herd behavior and aggregate fluctuations in financial markets. *Macroeconomic dynamics*, 4(02):170–196, 2000.
- [3] David Hirshleifer and Siew Hong Teoh. Herd behaviour and cascading in capital markets: A review and synthesis. *European Financial Management*, 9(1):25–66, 2003.
- [4] Jiaqi Ge. Who creates housing bubbles? an agent-based study. In *International Workshop on Multi-Agent Systems and Agent-Based Simulation*, pages 143–150. Springer, 2013.
- [5] Wenli Li and Rui Yao. The life-cycle effects of house price changes. *Journal of Money, Credit and banking*, 39(6):1375–1409, 2007.
- [6] Fumio Hayashi and Edward C Prescott. The 1990s in japan: A lost decade. *Review of Economic Dynamics*, 5(1):206–235, 2002.
- [7] Robert J Shiller. *The subprime solution: How today’s global financial crisis happened, and what to do about it*. Princeton University Press, 2012.
- [8] Robert J Shiller. *Understanding recent trends in house prices and home ownership*. National Bureau of Economic Research, 2007.
- [9] Frederic S Mishkin. Over the cliff: From the subprime to the global financial crisis. *The Journal of Economic Perspectives*, 25(1):49–70, 2011.

- [10] Jose A Scheinkman and Wei Xiong. Overconfidence and speculative bubbles. *Journal of political Economy*, 111(6):1183–1220, 2003.
- [11] Margaret Hwang Smith and Gary Smith. Bubble, bubble, where’s the housing bubble? *Brookings Papers on Economic Activity*, 2006(1):1–67, 2006.
- [12] Karl E Case and Robert J Shiller. Is there a bubble in the housing market? *Brookings Papers on Economic Activity*, 2003(2):299–362, 2003.
- [13] Karl E Case and Robert J Shiller. *The behavior of home buyers in boom and post-boom markets*. National Bureau of Economic Research Cambridge, Mass., USA, 1988.
- [14] Edward L Glaeser, Joseph Gyourko, and Albert Saiz. Housing supply and housing bubbles. *Journal of urban Economics*, 64(2):198–217, 2008.
- [15] Allen C Goodman and Thomas G Thibodeau. Where are the speculative bubbles in US housing markets? *Journal of Housing Economics*, 17(2):117–137, 2008.
- [16] Charles Himmelberg, Christopher Mayer, and Todd Sinai. Assessing high house prices: Bubbles, fundamentals and misperceptions. *The Journal of Economic Perspectives*, 19(4):67–92, 2005.
- [17] Jonathan McCarthy and Richard W Peach. Are home prices the next bubble? *Economic Policy Review*, 10(3), Dec 2004.
- [18] John H Cochrane. *Asset Pricing:(Revised Edition)*. Princeton university press, 2009.
- [19] Matthew S Yiu, Jun Yu, and Lu Jin. Detecting bubbles in Hong Kong residential property market. *Journal of Asian Economics*, 28:115–124, 2013.
- [20] Gaia. Garino and Lucio Sarno. Speculative bubbles in U.K. house prices: Some new evidence. *Southern Economic Journal*, 70(4):777–795, 2004.

- [21] Vyacheslav Mikhed and Petr Zemčík. Testing for bubbles in housing markets: A panel data approach. *The Journal of Real Estate Finance and Economics*, 38(4):366–386, 2007.
- [22] Bala Arshanapalli and William Nelson. A cointegration test to verify the housing bubble. *The International Journal of Business and Finance Research*, 2(2):35–43, 2008.
- [23] Thomas Lux and Frank Westerhoff. Economics crisis. *Nature Physics*, 5(1):2–3, 2009.
- [24] Wei-Xing Zhou and Didier Sornette. 2000–2003 real estate bubble in the UK but not in the USA. *Physica A: Statistical Mechanics and its Applications*, 329(1):249–263, 2003.
- [25] Louis Bachelier. *Louis Bachelier’s theory of speculation: the origins of modern finance*. Princeton University Press, 2011.
- [26] Mark Buchanan. What has econophysics ever done for us? *Nature Physics*, 9(6):317, 2013.
- [27] Albert Einstein. *Investigations on the Theory of the Brownian Movement*. Courier Dover Publications, 1956.
- [28] Thomas Lux and Michele Marchesi. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature*, 397(6719):498–500, 1999.
- [29] Rosario N Mantegna and H Eugene Stanley. Scaling behaviour in the dynamics of an economic index. *Nature*, 376(6535):46–49, 1995.
- [30] Xavier Gabaix, Parameswaran Gopikrishnan, Vasiliki Plerou, and H Eugene Stanley. A theory of power-law distributions in financial market fluctuations. *Nature*, 423(6937):267–270, 2003.
- [31] Stephen Malpezzi. Hedonic pricing models: a selective and applied review. *Section in Housing Economics and Public Policy: Essays in Honor of Duncan MacLennan*, 2003.

- [32] Robert Hill. Hedonic price indexes for housing. *OECD Statistics Working Papers*, 2011(1):0-1, 2011.
- [33] Esmeralda A Ramalho and JS Ramalho. Hedonic functions, hedonic methods, estimation methods and Dutot and Jevons house price indexes: are there any links. Technical report, Mimeo, Universidade de Évora, 2011.
- [34] Hsiao-Jung Teng, Chin-Oh Chang, and KW Chau. Housing bubbles: A tale of two cities. *Habitat International*, 39:8–15, 2013.
- [35] H Eugene Stanley, Viktor Afanasyev, Luis A Nunes Amaral, SV Buldyrev, AL Goldberger, Steve Havlin, Harry Leschhorn, P Maass, Rosario N Mantegna, C-K Peng, et al. Anomalous fluctuations in the dynamics of complex systems: from DNA and physiology to econophysics. *Physica A: Statistical Mechanics and its Applications*, 224(1-2):302–321, 1996.
- [36] Yaneer Bar-Yam. General features of complex systems. *Encyclopedia of Life Support Systems (EOLSS)*, UNESCO, EOLSS Publishers, Oxford, UK, 2002.
- [37] Jianxi Gao, Baruch Barzel, and Albert-László Barabási. Universal resilience patterns in complex networks. *Nature*, 530(7590):307, 2016.
- [38] Marten Scheffer, Jordi Bascompte, William A Brock, Victor Brovkin, Stephen R Carpenter, Vasilis Dakos, Hermann Held, Egbert H Van Nes, Max Rietkerk, and George Sugihara. Early-warning signals for critical transitions. *Nature*, 461(7260):53–59, 2009.
- [39] Tom Andersen, Jacob Carstensen, Emilio Hernandez-Garcia, and Carlos M Duarte. Ecological thresholds and regime shifts: approaches to identification. *Trends in Ecology & Evolution*, 24(1):49–57, 2009.
- [40] Marten Scheffer, Steve Carpenter, Jonathan A Foley, Carl Folke, and Brian Walker. Catastrophic shifts in ecosystems. *Nature*, 413(6856):591, 2001.
- [41] Marten Scheffer and Stephen R Carpenter. Catastrophic regime shifts in ecosystems: linking theory to observation. *Trends in ecology & evolution*, 18(12):648–656, 2003.

- [42] Vasilis Dakos, Marten Scheffer, Egbert H van Nes, Victor Brovkin, Vladimir Petoukhov, and Hermann Held. Slowing down as an early warning signal for abrupt climate change. *Proceedings of the National Academy of Sciences*, 105(38):14308–14312, 2008.
- [43] Marten Scheffer. Complex systems: foreseeing tipping points. *Nature*, 467(7314):411–412, 2010.
- [44] Thomas Kleinen, Hermann Held, and Gerhard Petschel-Held. The potential role of spectral properties in detecting thresholds in the earth system: application to the thermohaline circulation. *Ocean Dynamics*, 53(2):53–63, 2003.
- [45] Marten Scheffer, Stephen R Carpenter, Timothy M Lenton, Jordi Bascompte, William Brock, Vasilis Dakos, Johan Van de Koppel, Ingrid A Van de Leemput, Simon A Levin, Egbert H Van Nes, et al. Anticipating critical transitions. *Science*, 338(6105):344–348, 2012.
- [46] Egbert H Van Nes and Marten Scheffer. Slow recovery from perturbations as a generic indicator of a nearby catastrophic shift. *The American Naturalist*, 169(6):738–747, 2007.
- [47] SR Carpenter and WA Brock. Rising variance: a leading indicator of ecological transition. *Ecology letters*, 9(3):311–318, 2006.
- [48] Vishweshha Guttal and Ciriya Jayaprakash. Changing skewness: an early warning signal of regime shifts in ecosystems. *Ecology letters*, 11(5):450–460, 2008.
- [49] James Peng Lung Tan and Siew Ann Cheong. Critical slowing down associated with regime shifts in the US housing market. *The European Physical Journal B*, 87(2):1–10, 2014.
- [50] Wei-Xing Zhou and Didier Sornette. Analysis of the real estate market in Las Vegas: Bubble, seasonal patterns, and prediction of the CSW indices. *Physica A: Statistical Mechanics and its Applications*, 387(1):243–260, 2008.

- [51] Darrell Jiajie Tay, Chung-I Chou, Sai-Ping Li, Shang You Tee, and Siew Ann Cheong. Bubbles are departures from equilibrium housing markets: Evidence from Singapore and Taiwan. *PloS one*, 11(11):e0166004, 2016.
- [52] James Crotty. Structural causes of the global financial crisis: a critical assessment of the new financial architecture. *Cambridge journal of economics*, 33(4):563–580, 2009.
- [53] Dennis R Capozza, Patric H Hendershott, and Charlotte Mack. An anatomy of price dynamics in illiquid markets: analysis and evidence from local housing markets. *Real Estate Economics*, 32(1):1–32, 2004.
- [54] Jesse Colombo. It’s not a bubble until it’s officially denied, Malaysia edition. <http://www.forbes.com/sites/jessecolombo/2013/10/18/its-not-a-bubble-until-its-officially-denied-malaysia-edition/>, Oct 2013. Accessed 2015 Nov 17.
- [55] Jesse Colombo. It’s not a bubble until it’s officially denied, Singapore edition. <http://www.forbes.com/sites/jessecolombo/2014/01/16/its-not-a-bubble-until-its-officially-denied-singapore-edition/>, Jan 2014. Accessed 2015 Nov 17.
- [56] Jesse Colombo. Here’s what The Philippines bubble deniers are getting wrong. <http://www.forbes.com/sites/jessecolombo/2013/11/28/heres-what-the-philippine-bubble-deniers-are-getting-wrong/>, Nov 2013. Accessed 2015 Nov 17.
- [57] Hao Meng, Wen-Jie Xie, Zhi-Qiang Jiang, Boris Podobnik, Wei-Xing Zhou, and H Eugene Stanley. Systemic risk and spatiotemporal dynamics of the US housing market. *Scientific reports*, 4:3655, 2014.
- [58] Wei-Xing Zhou and Didier Sornette. Is there a real-estate bubble in the US? *Physica A: Statistical Mechanics and its Applications*, 361(1):297–308, 2006.
- [59] Diego Ardila, Peter Cauwels, Dorsa Sanadgol, and Didier Sornette. Is there a real estate bubble in Switzerland? *arXiv preprint arXiv:1303.4514*, 2013.

- [60] Takaaki Ohnishi, Takayuki Mizuno, Chihiro Shimizu, and Tsutomu Watanabe. Power laws in real estate prices during bubble periods. *International Journal of Modern Physics: Conference Series*, 16:61–81, 2012.
- [61] Chung-I Chou and SP Li. House price distributions of Taiwan: a preliminary study. *arXiv preprint arXiv:1008.1376*, 2010.
- [62] Niall MacKay. London house prices are power-law distributed. *arXiv preprint arXiv:1012.3039*, 2010.
- [63] Ngai Ming Yip and Chin-Oh Chang. Housing in Taiwan: state intervention in a market driven housing system. *The journal of comparative Asian development*, 2(1):93–113, 2003.
- [64] Urban Redevelopment Authority. Real estate information system (realis) [online]. <https://spring.ura.gov.sg/lad/ore/login/findOutMore.cfm>, Jan 2013. Accessed 2016-01-13.
- [65] Government of Singapore. Resale flat prices [online]. <https://ref.data.gov.sg/Metadata/SGMetadata.aspx?id=2315050000000014700J&mid=171026&t=TEXTUAL>, 2014. Accessed 2016-01-13.
- [66] Department of Land Administration. Real estate transaction file (in traditional chinese) [online]. <http://plvr.land.moi.gov.tw/DownloadOpenData>, Mar 2016.
- [67] Department of Statistics, Singapore. Consumer price index, base year 2014 = 100, annual [online]. <http://www.singstat.gov.sg/statistics/browse-by-theme/prices>, Dec 2015.
- [68] National Statistics, Republic of China. Annual change of consumer price indices [online]. <http://eng.stat.gov.tw/ct.asp?xItem=12092&ctNode=1558>, Dec 2015.
- [69] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

- [70] VF Pisarenko and D Sornette. Robust statistical tests of dragon-kings beyond power law distributions. *The European Physical Journal-Special Topics*, 205(1):95–115, 2012.
- [71] Didier Sornette. Dragon-kings, black swans and the prediction of crises. *Social Science Research Network*, (09-36), 2009.
- [72] Liang Jiang, Peter CB Phillips, and Jun Yu. A new hedonic regression for real estate prices applied to the Singapore residential market. *Social Science Research Network*, 2014.
- [73] Hao Meng, Wen-Jie Xie, and Wei-Xing Zhou. Club convergence of house prices: Evidence from China’s ten key cities. *International Journal of Modern Physics B*, 29(24):1550181, 2015.
- [74] Singapore Land Authority. Speech by mr mah bow tan, minister for national development, on policy changes affecting the property market, at parliamentary sitting [online]. <http://www.sla.gov.sg/News/tabid/142/articleid/241/category/Press\%20Releases/parentId/97/year/2005/Default.aspx>, Jul 2005.
- [75] Adrian Drăgulescu and Victor M Yakovenko. Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States. *Physica A: Statistical Mechanics and its Applications*, 299(1):213–221, 2001.
- [76] VictorM. Yakovenko and A.Christian Silva. *Two-class Structure of Income Distribution in the USA: Exponential Bulk and Power-law Tail*. New Economic Windows. Springer Milan, 2005.
- [77] Yoshi Fujiwara, Wataru Souma, Hideaki Aoyama, Taisei Kaizoji, and Masanao Aoki. Growth and fluctuations of personal income. *Physica A: Statistical Mechanics and its Applications*, 321(3):598–604, 2003.

- [78] A. Chatterjee and B. K. Chakrabarti. Kinetic exchange models for income and wealth distributions. *The European Physical Journal B*, 60(2):135–149, 2007.
- [79] Jean-Philippe Bouchaud and Marc Mezard. Wealth condensation in a simple model of economy. *Physica A: Statistical Mechanics and its Applications*, 282(34):536 – 545, 2000.
- [80] Forbes. Singapore 50 richest 2015 ranking [online]. <http://www.forbes.com/singapore-billionaires/list/>, 2015. Accessed 2016-01-13.
- [81] Accounting The Directorate General of Budget and Executive Yuan Statistics (DGBAS). Household income findings [in chinese]. <http://win.dgbas.gov.tw/fies/index.asp>, 2012. Accessed 06 June 2016.
- [82] Academia Sinica Center for Survey Research, RCHSS. Survey of family income and expenditure. https://srda.sinica.edu.tw/gov/group_en/18, 2012. Accessed 06 June 2016.
- [83] Forbes. Taiwan 50 richest 2015 ranking [online]. <http://www.forbes.com/taiwan-billionaires/list/>, 2015. Accessed 2016-01-13.
- [84] Chung-Han Chan, Yih-Min Wu, Tai-Lin Tseng, Ting-Li Lin, and Chien-Chih Chen. Spatial and temporal evolution of b-values before large earthquakes in Taiwan. *Tectonophysics*, 532:215–222, 2012.
- [85] Nicolaas W Posthumus. The tulip mania in holland in the years 1636 and 1637. *Journal of Economic and Business History*, 1(4), 1929.
- [86] Peter M Garber. Famous first bubbles. *The Journal of Economic Perspectives*, 4(2):35–54, 1990.
- [87] Morgan Housel. Bubble, bubble, where’s the bubble? <https://www.usatoday.com/story/money/markets/2013/12/09/bubble-talk-overblown/3923375/>, Dec 2013. Accessed 2015 Nov 17.

- [88] Diego Ardila, Dorsa Sanadgol, Peter Cauwels, and Didier Sornette. Identification and critical time forecasting of real estate bubbles in the usa. *Quantitative Finance*, 17(4):613–631, 2017.
- [89] Zhi-Qiang Jiang, Wei-Xing Zhou, Didier Sornette, Ryan Woodard, Ken Bastiaensen, and Peter Cauwels. Bubble diagnosis and prediction of the 2005–2007 and 2008–2009 chinese stock market bubbles. *Journal of economic behavior & organization*, 74(3):149–162, 2010.
- [90] Annelies J Veraart, Elisabeth J Faassen, Vasilis Dakos, Egbert H van Nes, Miquel Lüring, and Marten Scheffer. Recovery rates reflect distance to a tipping point in a living system. *Nature*, 481(7381):357, 2012.
- [91] Ingrid A van de Leemput, Marieke Wichers, Angélique OJ Cramer, Denny Borsboom, Francis Tuerlinckx, Peter Kuppens, Egbert H van Nes, Wolfgang Viechtbauer, Erik J Giltay, Steven H Aggen, et al. Critical slowing down as early warning for the onset and termination of depression. *Proceedings of the National Academy of Sciences*, 111(1):87–92, 2014.
- [92] Christopher Carr, Ralf Ritter, CG Wade, Charles S Adams, and Kevin J Weatherill. Nonequilibrium phase transition in a dilute rydberg ensemble. *Physical review letters*, 111(11):113901, 2013.
- [93] James Tan and Siew Ann Cheong. The regime shift associated with the 2004–2008 US housing market bubble. *PloS one*, 11(9):e0162140, 2016.
- [94] Vasilis Dakos, Egbert H van Nes, Raúl Donangelo, Hugo Fort, and Marten Scheffer. Spatial correlation as leading indicator of catastrophic shifts. *Theoretical Ecology*, 3(3):163–174, 2010.
- [95] Tom Nunlist. China’s housing market: The iron bubble. <http://knowledge.ckgsb.edu.cn/2017/07/31/real-estate/china-housing-market-iron-bubble/>, Jul 2017. Accessed 2017 Aug 21.

- [96] David Scutt. Nomura: China's housing bubble is spreading to its smallest cities. <http://www.businessinsider.com/china-housing-bubble-spreading-to-smaller-cities-2017-6/?IR=T>, Jun 2017. Accessed 2017 Aug 21.
- [97] Enda Curran. dangerous situation for hong kong property market, says finance chief. <https://www.bloomberg.com/news/articles/2017-06-19/-dangerous-situation-for-hong-kong-property-market-says-chan>, Jun 2017. Accessed 2017 Aug 21.
- [98] Venus Wu and Clare Jim. Hong kong unveils fresh property curbs as chinese buyers rush in. <http://www.reuters.com/article/us-hongkong-property-measures-idUSKBN12Z13L?il=0>, Nov 2016. Accessed 2017 Aug 19.
- [99] Frederik Balfour. Chinese buyers returning to hong kong housing, wheelock says. <https://www.bloomberg.com/news/articles/2016-10-04/chinese-buyers-returning-to-hong-kong-housing-wheelock-says>, Oct 2016. Accessed 2017 Aug 19.
- [100] Norden E Huang, Zheng Shen, Steven R Long, Manli C Wu, Hsing H Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H Liu. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. In *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences*, volume 454, pages 903–995. The Royal Society, 1998.
- [101] Zhaohua Wu, Norden E Huang, Steven R Long, and Chung-Kang Peng. On the trend, detrending, and variability of nonlinear and nonstationary time series. *Proceedings of the National Academy of Sciences*, 104(38):14889–14894, 2007.
- [102] Abdel-Ouahab Boudraa and Jean-Christophe Cexus. Emd-based signal filtering. *IEEE Transactions on Instrumentation and Measurement*, 56(6):2196–2202, 2007.

- [103] Lennart Ljung. System identification. In *Signal analysis and prediction*, pages 163–173. Springer, 1998.
- [104] Norden Eh Huang. *Hilbert-Huang transform and its applications*, volume 16. World Scientific, 2014.
- [105] A Grossmann, Richard Kronland-Martinet, and J Morlet. Reading and understanding continuous wavelet transforms. In *Wavelets*, pages 2–20. Springer, 1990.
- [106] Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.
- [107] James B Ramsey. Wavelets in economics and finance: Past and future. *Studies in Nonlinear Dynamics & Econometrics*, 6(3), 2002.
- [108] Satish Sinha, Partha S Routh, Phil D Anno, and John P Castagna. Spectral decomposition of seismic data with continuous-wavelet transform. *Geophysics*, 70(6):P19–P25, 2005.
- [109] Yazhen Wang. Jump and sharp cusp detection by wavelets. *Biometrika*, 82(2):385–397, 1995.
- [110] C Cattani. Haar wavelet-based technique for sharp jumps classification. *Mathematical and Computer Modelling*, 39(2-3):255–278, 2004.
- [111] James Ze Wang, Gio Wiederhold, Oscar Firschein, and Sha Xin Wei. Content-based image indexing and searching using daubechies’ wavelets. *International Journal on Digital Libraries*, 1(4):311–328, 1998.
- [112] Chris A Boulton, Lesley C Allison, and Timothy M Lenton. Early warning signals of atlantic meridional overturning circulation collapse in a fully coupled climate model. *Nature communications*, 5:5752, 2014.
- [113] M. Kac, J. Kiefer, and J. Wolfowitz. On tests of normality and other tests of goodness of fit based on distance methods. *Ann. Math. Statist.*, 26(2):189–211, 06 1955.

- [114] Ralph B D’Agostino. Transformation to normality of the null distribution of g_1 . *Biometrika*, pages 679–681, 1970.
- [115] Carlos M Jarque and Anil K Bera. A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique*, pages 163–172, 1987.
- [116] SS Shaphiro and MB Wilk. An analysis of variance test for normality. *Biometrika*, 52(3):591–611, 1965.
- [117] T. W. Anderson and D. A. Darling. Asymptotic theory of certain ”goodness of fit” criteria based on stochastic processes. *Ann. Math. Statist.*, 23(2):193–212, 06 1952.
- [118] T. W. Anderson and D. A. Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954.
- [119] Frank J. Massey Jr. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.
- [120] Hubert W. Lilliefors. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402, 1967.
- [121] Nornadiiah Mohd Razali, Yap Bee Wah, et al. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of statistical modeling and analytics*, 2(1):21–33, 2011.
- [122] Mark EJ Newman. Power laws, Pareto distributions and Zipf’s law. *Contemporary physics*, 46(5):323–351, 2005.
- [123] Vasiliki Plerou, Parameswaran Gopikrishnan, Luis A Nunes Amaral, Martin Meyer, and H Eugene Stanley. Scaling of the distribution of price fluctuations of individual companies. *Physical review e*, 60(6):6519, 1999.

- [124] Parameswaran Gopikrishnan, Vasiliki Plerou, Luis A Nunes Amaral, Martin Meyer, and H Eugene Stanley. Scaling of the distribution of fluctuations of financial market indices. *Physical Review E*, 60(5):5305, 1999.
- [125] Boon Kin Teh and Siew Ann Cheong. The asian correction can be quantitatively forecasted using a statistical model of fusion-fission processes. *PloS one*, 11(10):e0163842, 2016.
- [126] George K Zipf. Human behavior and the principle of least effort. *Ed: Addison-Wesley, Reading, MA*, 1949.
- [127] Ramon Ferrer i Cancho and Richard V. Solé. The small world of human language. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1482):2261–2265, 2001.
- [128] Felix Auerbach. Das gesetz der bevölkerungskonzentration. *Petermanns Geographische Mitteilungen*, 59:74–76, 1913.
- [129] Xavier Gabaix and Yannis M Ioannides. The evolution of city size distributions. *Handbook of regional and urban economics*, 4:2341–2378, 2004.
- [130] Benoit Mandelbrot. The Pareto-Levy law and the distribution of income. *International Economic Review*, 1(2):79–106, 1960.
- [131] Victor M Yakovenko and J Barkley Rosser Jr. Colloquium: Statistical mechanics of money, wealth, and income. *Reviews of Modern Physics*, 81(4):1703, 2009.
- [132] Michal Brzezinski. Do wealth distributions follow power laws? evidence from rich lists. *Physica A: Statistical Mechanics and its Applications*, 406:155–162, 2014.
- [133] Ronald Aylmer Fisher. On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, 41:155–160, 1912.

- [134] Bungon Kumphon. Maximum entropy and maximum likelihood estimation for the three-parameter kappa distribution. *Open Journal of Statistics*, 2(4):415–419, 2012.
- [135] Z Hradil and J Reháček. Likelihood and entropy for statistical inversion. In *Journal of Physics: Conference Series*, volume 36, page 55. IOP Publishing, 2006.
- [136] Hirotogu Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*. Springer New York, New York, NY, 1998.
- [137] Douglas M Bates and Donald G Watts. *Nonlinear regression analysis and its applications*. Wiley, 1988.
- [138] Jeffrey M. Wooldridge. Applications of generalized method of moments estimation. *The Journal of Economic Perspectives*, 15(4):87–100, 2001.
- [139] A. Colin Cameron and Frank A.G. Windmeijer. An r-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, 77(2):329 – 342, 1997.
- [140] J. Janczura and R. Weron. Black swans or dragon-kings? a simple test for deviations from the power law. *The European Physical Journal Special Topics*, 205(1):79–93, 2012.
- [141] American Statistical Association. ASA p-value statement viewed > 150,000 times. <https://www.amstat.org/ASA/News/ASA-P-Value-Statement-Viewed-150000-Times.aspx>, 2016. Accessed: 2017-03-07.
- [142] Ronald L Wasserstein and Nicole A Lazar. The asa’s statement on p-values: context, process, and purpose. *Am Stat*, 70(2):129–133, 2016.
- [143] Monya Baker. Statisticians issue warning over misuse of p values. *Nature*, 531:151, 2016.

- [144] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one*, 9(1):e85777, 2014.
- [145] Shan Yu, Andreas Klaus, Hongdian Yang, and Dietmar Plenz. Scale-invariant neuronal avalanche dynamics and the cut-off in size distributions. *PloS one*, 9(6):e99761, 2014.
- [146] Najja Marshall, Nicholas M Timme, Nicholas Bennett, Monica Ripp, Edward Lautzenhiser, and John M Beggs. Analysis of power laws, shape collapses, and neural complexity: new techniques and matlab support via the ncc toolbox. *Frontiers in Physiology*, 7, 2016.
- [147] Edwin JG Pitman and Edwin James George Pitman. *Some basic theory for statistical inference*, volume 7. Chapman and Hall London, 1979.
- [148] Govindan Rangarajan and Mingzhou Ding. *Processes with long-range correlations: Theory and applications*, volume 621. Springer Science & Business Media, 2003.
- [149] Housing Development Board. Buying a flat. <http://www.hdb.gov.sg/cs/infoweb/residential/buying-a-flat>, 2005. Accessed Accessed 01 Mar 2016.
- [150] Singapore Land Authority. Foreign ownership of properties [online]. <http://www.sla.gov.sg/Services/RestrictiononForeignOwnershipofLandedProperty.aspx>, 2005. Accessed 30 May 2016.
- [151] Luc Devroye. *Non-uniform random variate generation*. Springer, 1 edition, 1986.
- [152] John E Angus. The probability integral transform and related results. *SIAM review*, 36(4):652–654, 1994.
- [153] Ronald Pyke. Spacings. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 395–449, 1965.

Appendix A

Housing in Singapore and Taiwan

A.1 Housing Types in Singapore and Taiwan

There are 3 main housing types in Singapore, namely, (i) the Housing Development Board (HDB) flats, (ii) the condominium apartments and (iii) landed properties. HDB flats are typically high-rise public housing which majority of the average Singaporean live in. Condominium apartments cater more to the upper middle class and have common facilities such as swimming pools, gymnasiums and entertainment rooms that the residents have access to. The landed properties are properties that are detached or semi-detached terraces and bungalows. These properties also usually have a small front/back yard and porch attached.

Generally, the Singapore government provides subsidies to first-time Singaporean home owners who purchase HDB flats and executive condominiums (ECs) apartments. The HDBs is a public housing scheme which is strictly regulated by the Singapore government. Only Singaporeans and Singapore Permanent Residents (SPRs) are allowed to purchase HDB flats. With effect from August 2013, SPR must wait for 3 years after obtaining SPR status before purchasing HDB flats. For most new HDB flat purchases, there is a minimum occupancy period for which the owners cannot own other HDB flat or private property in Singapore or overseas [149].

Condominiums and landed properties are collectively known as private properties in Singapore. This is with the exception of new ECs that have not fulfilled the minimum occupancy period. The condominiums (including ECs that have passed

the minimum occupation period) can be freely bought and sold with very minimal restrictions. The landed properties in mainland Singapore however, cannot be owned by foreigners. Foreigners can however, own landed properties on Sentosa island only if it is owner occupied [150].

In Taiwan, most housing properties are privately owned, with only a few exceptions of public housings owned by the government. When one buys a house, an apartment or a studio, he also owns part of the land permanently where the building is located. This is different from places like Hong Kong and China where one only owns the land for a certain period of time.

There are several housing types in Taiwan. Tou Tians are old Taiwanese style detached houses. These houses do not have front or back yards like the detached houses in the US and they have usually existed for at least several decades. There are a few newer types of detached houses or town houses that are built in recent years, e.g. in the suburbs of (Greater Taipei Area) GTA. The prices of Tou Tians vary a lot, depending on their locations. Tou Tians in the GTA are usually expensive even if they are very old, simply because of the land owned by the owner.

Although Zhu Zai Da Lou, Gong Yu and Hua Sha all have apartments and/or studios (Tao Fang), there are differences among them. Zhu Zai Da Lou are high rise buildings (usually much more expensive and newer) and many of them are located in communities that have good public facilities. In some cases, e.g., in Taipei, they belong to the tail of the power law in the distribution. Hua Sha are buildings with elevators (usually under 10 floors and newer). Gong Yus are usually buildings with 5 or fewer floors. They are older (20 years or more) and without any public facilities such as elevators. Hua Sha and Gong Yu are usually cheaper. Most middle class people can only afford to buy an apartment in Hua Sha or Gong Yu in recent years.

A.2 Year-by-Year plots of the Singapore Housing Market

The year-by-year plots for each housing types in Singapore are obtained by plotting the yearly contributions to the CDF. The yearly contribution to the HDB CDF

generally follows a straight line which suggests that it is in an equilibrium from the period 1995 to 2009. The second exponential regime only appears after 2009. For the condominium CDF contribution, we can see that the distribution is a stable exponential before 2007. A bump above the empirically fitted distribution only appears after the year 2007 and persists beyond that indicates a bubble forming in the condominium segment of the Singapore housing market. The landed properties is a straight line in the log-log plots for all the years of the data set which shows that the landed properties in Singapore is power law distributed and stable throughout the period 1995 to 2014.

Figure A.1: HDB CDF Year By Year Plots from 2000 to 2012

Figure A.2: Condominium CDF Year By Year Plots from 1995 to 2014

Figure A.3: Landed properties CDF Year By Year Plots from 1995 to 2014

A.3 Spatial and Temporal Heat Map of the Singapore Housing Market

The price of the various housing types can be studied in space and time simultaneously using heat maps. To obtain the heat maps, we store the monthly upper-quartile price for each sector. Using the colormap function (with "jet" color scheme) in MATLAB ©, we set the lowest and highest price observed in the data set for each of the housing types to represent blue and red respectively. Next, each of the upper quartile price stored is transferred into a color between red and blue using linear regression. Each frame in the heat map movie then shows the upper quartile price of the different sectors, for a particular housing type. Here, we show selected frames for each of the different housing types for the purpose of illustrating the spatial-temporal price dynamics in the Singapore housing market. For the full duration of the data set, please refer to the attached videos in the Supporting Information in [51].

From the heat map, we can see that the price increase started with a spike in sector 2. This price increase then spreads to the neighboring districts in the following months. This general dynamics applies to all three different housing types. From the heat map, we are also able to detect the bubble emerging in the condominium segment of the Singapore housing market. For the HDB flats and landed properties, an increase in price is quickly dissipated into the other sectors. On the other hand, we can evidently see that the condominiums in sector 2 was much higher priced than the other sectors during the peak of the bubble in September 2007.

Figure A.4: HDB Heat Map from 2000 to 2002

Figure A.5: HDB Heat Map from 2003 to 2005

Figure A.6: HDB Heat Map from 2006 to 2008

Figure A.7: HDB Heat Map from 2009 to 2010

Figure A.8: HDB Heat Map from 2011 to Feb 2012

Figure A.9: Condominium apartment Heat Map from 1995 to 1997

Figure A.10: Condominium apartment Heat Map from 1998 to 2000

Figure A.11: Condominium apartment Heat Map from 2001 to 2003

Figure A.12: Condominium apartment Heat Map from 2004 to 2006

Figure A.13: Condominium apartment Heat Map from 2007 to 2009

Figure A.14: Condominium apartment Heat Map from 2010 to 2012

Figure A.15: Condominium apartment Heat Map from 2013 to 2014

Figure A.16: Landed properties Heat Map from 1995 to 1997

Figure A.17: Landed properties Heat Map from 1998 to 2000

Figure A.18: Landed properties Heat Map from 2001 to 2003

Figure A.19: Landed properties Heat Map from 2004 to 2006

Figure A.20: Landed properties Heat Map from 2007 to 2009

Figure A.21: Landed properties Heat Map from 2010 to 2012

Figure A.22: Landed properties Heat Map from 2013 to 2014

Appendix B

Consumer Price Indexes

The consumer price index in Singapore is maintain by the Singapore Department of Statistics. Although the CPI is reported each month, in this thesis, we use an aggregated yearly CPI for each year, which is computed using the simple average of 12 monthly CPI. The CPI in Singapore is updated regularly every five years. Here, we use the CPI with 2014 as the base year and show the components of the CPI in Table B.1, as well as the CPI for the period 1995 to 2014 in Figure B.1.

The consumer price index of Taiwan is maintained by the Directorate-General of Budget, Accounting and Statistics. The CPI is reported each month, but we use an aggregated yearly CPI for each year, which is computed using the simple average of 12 monthly CPI. Here, we use the latest CPI update, which uses 2011 as the base year. The components of the CPI in Table B.2. Since we only use Taiwan housing data from 2012 to Oct 2014, we present the CPI for this period, **2012: 101.93, 2013: 102.80, 2014: 104.02.**

The consumer price index in Hong Kong (HK) is maintain by the Census and Statistics Department. Similarly, the CPI is posted reported month, but we use an aggregated yearly CPI for each year, which is computed using the simple average of 12 monthly CPI. The CPI in HK is updated regularly every five years. Here, we use the CPI with 2015 as the base year and show the components of the CPI in Table B.3, as well as the CPI for the period 1992 to 2011 in Figure B.2.

Each of the raw housing price obtained from the sources are discounted by their respective CPIs, in order to reflect its value in the base year. We use the discount

formula,

$$X_{disc} = \frac{X_{raw} \times 100}{CPI}, \quad (\text{B.1})$$

where X_{disc} and X_{raw} is the discounted and raw price respectively. The CPI for Singapore and HK for each year can be obtained from Figures B.1 and B.2 respectively, while the CPI of Taiwan is given two paragraphs above (in bold).

Table B.1: The components of the CPI in Singapore. The base year for this index is 2014.

Component	Weight (in %)
Food	21.7
Clothing and Footwear	2.7
Housing and Utilities	26.3
Household Durables and Services	4.7
Health Care	6.1
Transport	15.8
Communications	3.9
Recreation and Culture	7.9
Education	6.1
Misc Goods and Services	4.8

Figure B.1: [Left] The Singapore CPI components. [Right] The CPI of Singapore from 1995 to 2014.

Table B.2: The components of the CPI in Taiwan. The base year for this index 2011.

Component	Weight (in %)
Food	25.19
Clothing	3.77
Housing	27.14
Transportation and Communication	15.34
Medical and Sports	4.92
Education and Leisure	16.84
Miscellaneous	6.8

Table B.3: The components of the CPI in HK. The base year for this index 2015.

Component	Weight (in %)
Misc Services	15.81
Transport	7.98
Misc Goods	3.56
Durables Goods	4.65
Clothings and Footwear	3.21
Alcohol, Drinks and Tobacco	0.54
Electricity, Gas and Water	2.67
Housing	34.29
Food	27.29

Figure B.2: [Left] The HK CPI components. [Right] The CPI of HK from 1992 to 2011.

Appendix C

Properties of $U(0, 1)$ and its Ordered Statistics

C.1 The Probability Integral transform

The probability integral transform of a random sample set X distributed by a known probability density function (PDF) $f_X(x)$, maps the set X to the standard uniform distribution, $U(0, 1)$ and can be written as,

$$F_X(X) = U, \quad (\text{C.1})$$

where $F(x)$ is the inverse function of the cumulative distribution function (CDF) of X .

The detailed and generalizable proof of (C.1) can be found in [151, 152], so we will only highlight the main idea (weak proof) here.

Assuming we know the probability density $f_X(x)$ of the set X , we obtain its CDF,

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{\infty} f_X(x') dx'. \quad (\text{C.2})$$

We introduce the substitution $F_X(X) = Y$ so we can write (C.2) as,

$$\begin{aligned}
F_Y(y) &= P(F_X(X) \leq y) \\
&= P(X \leq F_X^{-1}(y)) \\
&= F_X(F_X^{-1}(y)) \\
&= y \quad , \tag{C.3}
\end{aligned}$$

which is just the definition of the standard uniform random variable in the interval $[0, 1]$. In relation to our substitution $Y = F_X(X)$, putting X back into its own distribution gives $U(0, 1)$, which completes the proof. Note that in (C.3) we have assumed, that both the CDF and its inverse F_X^{-1} exists and is bijective, so $F_X(x)$ needs to be strictly increasing and continuous. The general case on this issue can be proofed using quantiles which is illustrated in [152].

Using the property (C.1), we can also invert to write,

$$X = F_X^{-1}(U), \tag{C.4}$$

which allows us to use the probability integral transform to generate other desired distributions from $U(0,1)$. This method is very convenient, as long as F_X^{-1} can be analytically written down.

C.2 Ranked Statistics of U(0,1)

For a set of observation $X = \{x_1, x_2, \dots, x_N\}$, where $x_1, x_2, \dots, x_N$ are independent and identically distributed (iid) by the density function $f_X(x)$, we define the *ordered set* of X as, $X_{ordered} = \{x_{(1)}, x_{(2)}, \dots, x_{(N)}\}$.

For any continuous $f(x)$, the probability of finding $X = x$ within a tolerance of $|\epsilon|$ is,

$$P(x - \epsilon \leq X \leq x + \epsilon) = P(X \in [x - \epsilon, x + \epsilon]) \approx 2\epsilon f(x), \quad (\text{C.5})$$

where the term on the RHS of the approximation represents the area of the rectangular strip under the density curve with length $f(x)$ and width 2ϵ .

We are interested in the k -th ranked iid, $x_{(k)}$, so we consider the probability (within a tolerance of ϵ) an observation is ranked k -th, which is,

$$2\epsilon f_k(x) \approx P(X_k \in [x - \epsilon, x + \epsilon]), \quad (\text{C.6})$$

where $f_k(x)$ is the order statistics of the distribution $f(x)$, which is interpreted as the probability of the k -th element found within the vicinity of a given value $x \pm \epsilon$, with vanishing ϵ . The RHS of (C.6) is equivalent to finding the ‘macrostate’ where strictly one x_i within $x \pm \epsilon$, strictly $k - 1$ of them lesser than $x - \epsilon$ and the rest (of the $N - k$) of them more than $x + \epsilon$. This is equivalently,

$$\begin{aligned} P(X_k \in [x - \epsilon, x + \epsilon]) &= \binom{N}{N-k} \binom{k}{k-1} 2\epsilon f(x) [F(x - \epsilon)]^{k-1} [1 - F(x + \epsilon)]^{N-k} \\ &= \frac{N!}{(k-1)!(N-k)!} 2\epsilon f(x) [F(x - \epsilon)]^{k-1} [1 - F(x + \epsilon)]^{N-k}, \end{aligned} \quad (\text{C.7})$$

where the combinatorics pre-factor accounts for the ‘microstates’. Combining the equations (C.6) and (C.7) and noting that the equality in (C.6) holds when $\epsilon \rightarrow 0$, we get,

$$f_k(x) = \frac{N!}{(k-1)!(N-k)!} f(x) [F(x)]^{k-1} [1 - F(x)]^{N-k}. \quad (\text{C.8})$$

For the $U(0, 1)$ distribution, its PDF is $f(x) = 1$ and CDF is $F(x) = x$, equation (C.8) becomes,

$$f_k(x) = \frac{N!}{(k-1)!(N-k)!} [F(x)]^{k-1} [1 - F(x)]^{N-k}. \quad (\text{C.9})$$

This is the order statistics of $U(0, 1)$

Appendix D

Finite Sample Effects (FSE) Corrections

D.1 Mathematical Derivations Related to Power-Law and Exponential Distributions

The following are the notations use in Chapter 4:

Table D.1: Notations used in Chapter 4: Mathematical derivations related to Power Law (PL) and Exponential Distribution (EXP)

Notation	Description	Example: PL, EXP
$f_*(x)$	Probability density function (PDF) for * distribution	$f_{PL}(x), f_{EXP}(x)$
$F_*(x)$	Cumulative density function (CDF) for * distribution	$F_{PL}(x), F_{EXP}(x)$
$F_*^{-1}(u)$	Inverse cumulative density function for * distribution	$F_{PL}^{-1}(u), F_{EXP}^{-1}(u)$
$\mathbb{L}_*$	Likelihood function for * distribution	$\mathbb{L}_{PL}, \mathbb{L}_{EXP}$
$\hat{P}^*$	Estimated parameter(s) for * distribution	$\hat{\alpha}, \hat{\beta}$
P_T^*	True parameter(s) for * distribution	α_T, β_T
RE_*	Relative estimation error, $\frac{\sqrt{\langle(\hat{P}^* - P_T^*)^2\rangle}}{P_T^*}$	RE_{PL}, RE_{EXP}
x_{max}	Postulated largest element, $x_{max} = F^{-1}(1 - \delta)$	$X_{max}^{PL}, X_{max}^{EXP}$

D.2 Adjustment for Finite Largest Element

In Chapter 4, we have demonstrated how the finite largest element effect (FLE) causes us to underestimate the value of $\langle \ln(x) \rangle$ (for PL) and $\langle x \rangle$ (for EXP). This results in an overestimated $\hat{\alpha}$ (for PL) and $\hat{\beta}$ (for EXP). We correct for this by adding back the truncated parts (from x_{max} to ∞) into $\langle \ln(x) \rangle$ and $\langle x \rangle$,

$$\begin{aligned} \left\langle \ln \left(\frac{x}{x_{min}} \right) \right\rangle_{adj} &= \left\langle \ln \left(\frac{x}{x_{min}} \right) \right\rangle_{data} \int_{x_{min}}^{x_{max}} f_{PL}(x) dx + \int_{x_{max}}^{\infty} \ln \left(\frac{x}{x_{min}} \right) f_{PL}(x) dx; \\ \langle x \rangle_{adj} &= \langle x \rangle_{data} \int_{x_{min}}^{x_{max}} f_{EXP}(x) dx + \int_{x_{max}}^{\infty} x f_{EXP}(x) dx. \end{aligned} \quad (D.1)$$

This yields the adjusted values,

$$\begin{aligned} \left\langle \ln \left(\frac{x}{x_{min}} \right) \right\rangle_{adj} &= \left\langle \ln \left(\frac{x}{x_{min}} \right) \right\rangle_{data} \left[1 - \left(\frac{x_{max}}{x_{min}} \right)^{1-\alpha} \right] + \\ &\quad \frac{1}{\alpha - 1} \left(\frac{x_{max}}{x_{min}} \right)^{1-\alpha} \left[(\alpha - 1) \ln \left(\frac{x_{max}}{x_{min}} \right) + 1 \right]; \quad (D.2) \\ \langle x \rangle_{adj} &= \langle x \rangle_{data} [1 - \exp(-\beta(x_{max} - x_{min}))] + \\ &\quad \frac{1}{\beta} (\beta x_{max} + 1) \exp(-\beta(x_{max} - x_{min})), \end{aligned}$$

for the PL and EXP distributions respectively (seperated by the semicolon). By inserting Eq. (D.2) into Eqs. (4.5) and (4.12) respectively, we obtain the nonlinear equations

$$\begin{aligned} &\left\langle \ln \left(\frac{x}{x_{min}} \right) \right\rangle_{data} + \frac{1}{1 - \hat{\alpha}_{adj}} + \\ &\quad \left(\frac{x_{max}}{x_{min}} \right)^{1-\hat{\alpha}_{adj}} \left[\ln \left(\frac{x_{max}}{x_{min}} \right) - \left\langle \ln \left(\frac{x}{x_{min}} \right) \right\rangle_{data} + \frac{1}{\hat{\alpha}_{adj} - 1} \right] = 0; \\ &\left[\hat{\beta}_{adj} (x_{max} - \langle x \rangle_{data}) + 1 \right] \exp \left[-\hat{\beta}_{adj} (x_{max} - x_{min}) \right] + \hat{\beta}_{adj} (\langle x \rangle_{data} - x_{min}) - 1 = 0 \end{aligned} \quad (D.3)$$

that can be solved to obtain $\hat{\alpha}_{adj}$ and $\hat{\beta}_{adj}$ respectively.

D.3 Theoretical Approximation of the Relative Error

To estimate the errors in P^* arising due to the FLE effect, x_{max} , we define the relative estimated parameter error as

$$RE = \frac{\sqrt{\left\langle \left(\hat{P} - P_T \right)^2 \right\rangle}}{P_T}. \quad (D.4)$$

The finite largest element x_{max} can be expressed in term of δ using Eqs. (4.8) and (4.16),

$$X_{max}^{PL} = x_{min} \delta^{\frac{1}{\alpha_T - 1}}; \quad X_{max}^{EXP} = x_{min} - \frac{\ln(\delta)}{\beta_T}, \quad (D.5)$$

the unadjusted terms $\langle \ln(x) - \ln(x_{min}) \rangle_{unadj}$ and $\langle x \rangle_{unadj}$ only contain $x \in [x_{min}, x_{max}]$, so they can be expressed as,

$$\left\langle \ln \left(\frac{x}{x_{min}} \right) \right\rangle_{unadj} = \int_{x_{min}}^{X_{max}^{PL}} \ln \left(\frac{x}{x_{min}} \right) f_{PL}(x|\alpha_T) dx; \quad \langle x \rangle_{unadj} = \int_{x_{min}}^{X_{max}^{EXP}} x f_{EXP}(x|\beta_T) dx. \quad (D.6)$$

Evaluating the integral for both equations gives,

$$\begin{aligned} \left\langle \ln \left(\frac{x}{x_{min}} \right) \right\rangle_{unadj} &= \frac{1}{\alpha_T - 1} \left\{ 1 - \left(\frac{X_{max}^{PL}}{x_{min}} \right)^{1 - \alpha_T} \left[(\alpha_T - 1) \ln \left(\frac{X_{max}^{PL}}{x_{min}} \right) + 1 \right] \right\}; \\ \langle x \rangle_{unadj} &= \frac{1}{\beta_T} \left\{ (\beta_T x_{min} + 1) - (\beta_T X_{max}^{EXP} + 1) \exp(-\beta_T (X_{max}^{EXP} - x_{min})) \right\}. \end{aligned} \quad (D.7)$$

To check the impact this has on the estimated parameters, we substitute the result in Eq. (D.7) into Eqs. (4.5) and (4.12) to obtain the unadjusted estimated parameter for PL distribution,

$$\begin{aligned} \hat{\alpha}_{unadj} &= 1 + \frac{\alpha_T - 1}{1 - \left(\frac{X_{max}^{PL}}{x_{min}} \right)^{1 - \alpha_T} \left[(\alpha_T - 1) \ln \left(\frac{X_{max}^{PL}}{x_{min}} \right) + 1 \right]} \\ &\approx \alpha_T + (\alpha_T - 1) \left(\frac{X_{max}^{PL}}{x_{min}} \right)^{1 - \alpha_T} \left[(\alpha_T - 1) \ln \left(\frac{X_{max}^{PL}}{x_{min}} \right) + 1 \right] + \\ &\quad \mathcal{O} \left\{ \left(\frac{X_{max}^{PL}}{x_{min}} \right)^{2(1 - \alpha_T)} \left[(\alpha_T - 1) \ln \left(\frac{X_{max}^{PL}}{x_{min}} \right) + 1 \right]^2 \right\} \end{aligned} \quad (D.8)$$

and for EXP distribution,

$$\begin{aligned}\hat{\beta}_{unadj} &= \frac{\beta_T}{1 - (\beta_T X_{max}^{EXP} + 1) \exp(-\beta_T (X_{max}^{EXP} - x_{min}))} \\ &\approx \beta_T + \beta_T [\beta_T X_{max}^{EXP} + 1] \exp(-\beta_T (X_{max}^{EXP} - x_{min})) + \\ &\quad \mathcal{O} \left\{ [\beta_T X_{max}^{EXP} + 1]^2 \exp(-2\beta_T (X_{max}^{EXP} - x_{min})) \right\}.\end{aligned}\tag{D.9}$$

The approximations in Eq. (D.8) and Eq. (D.9) are estimated using first order Taylor expansion. Rewriting $\hat{\alpha}_{unadj}$ and $\hat{\beta}_{unadj}$ in terms of δ from Eq. (D.5) will lead to

$$\begin{aligned}\alpha_{unadj} &\approx \alpha_T + (\alpha_T - 1) \delta [1 - \ln(\delta)] \\ \beta_{unadj} &\approx \beta_T + \beta_T \delta [1 - \ln(\delta) + \beta_T x_{min}].\end{aligned}\tag{D.10}$$

Rewriting the relative error for PL and EXP distribution in term of δ , we get

$$\begin{aligned}RE_{PL}(\alpha_{unadj}) &= \left(1 - \frac{1}{\alpha_T}\right) \delta [1 - \ln(\delta)], \quad \delta \in [0, 1] \\ RE_{EXP}(\beta_{unadj}) &= \delta [1 - \ln(\delta) + \beta_T x_{min}], \quad \delta \in [0, 1].\end{aligned}\tag{D.11}$$

While it is useful to note that both relative errors are monotonic increasing functions with $RE_{PL} \in [0, 1]$ and $RE_{EXP} \in [0, \beta_T x_{min}]$, it is more important to observe that for a given δ , RE_{PL} is always smaller or equal to RE_{EXP} . This sensitivity of the estimated parameter (relative error) on finite largest element ($x_{max}(\delta)$) provides us an insight on answering why it is easier for the PL distribution to pass significance testing compared to the EXP distribution using unadjusted parameters.

D.4 Relationship Between KS Distance, Distribution Noise, and Sample Size

For a reordered data set $X = \{x_1^{(em)}, x_2^{(em)}, \dots, x_N^{(em)}\}$ such that $x_i^{(em)} < x_j^{(em)}$ for $i < j$. If X is f_x distributed and has CDF F_x , the probability integral transform transforms X to the standard uniform distribution. Mathematically, this is the transformation (See Appendix C.1) $X = \{x_1^{(em)}, x_2^{(em)}, \dots, x_N^{(em)}\} \rightarrow \{F_x^{-1}(x_1^{(em)}) = u_1^{(em)}, F_x^{-1}(x_2^{(em)}) = u_2^{(em)}, \dots, F_x^{-1}(x_N^{(em)}) = u_N^{(em)}\}$. Thereafter, statistical significance testing is done by examining how plausible the empirically transformed set $U^{(em)} = \{u_1^{(em)}, u_2^{(em)}, \dots, u_N^{(em)}\} = U^{(em)}$ can be obtained by chance, from the standard uniform distribution, $U^{(s)} = \{u_1^{(s)}, u_2^{(s)}, \dots, u_N^{(s)}\}$. This can be done by comparing the KS distance and distribution noise value between $U^{(em)}$ and $U^{(s)}$. Since $U^{(em)}$ is fixed and depends only on the sample, we only need to focus on the samples drawn from standard uniform distribution ($U^{(s)}$) to understand the relationship between KS distance (d_{KS}), distribution noise (d_{DN}) and sample size (N).

We simulated 10^6 samples of various sizes N from $U(0, 1)$. Thus, for each N we have 10^6 simulated sample $U^{(s)} = \{u_1^{(s)}, u_1^{(s)}, \dots, u_N^{(s)}\}$ for $s = [1, 10^6]$ ordered such that $u_i^{(s)} \leq u_{i+1}^{(s)}$. For each sample then we calculate its KS distance d_{KS} as

$$d_{KS} = \forall_{i=1}^N \sup \left(\left| u_i - \frac{i}{N} \right| \right), \quad (\text{D.12})$$

and also its distribution noise d_{DN} using

$$\begin{aligned} d_{DN} &= \sqrt{\frac{\sum_{i=1}^N (u_i - u_{i-1})^2 [f(u_{i-1}, u_i) - 1]^2}{\sum_{i=1}^N (u_i - u_{i-1})^2}} \\ &= \sqrt{\frac{\sum_{i=1}^N (u_i - u_{i-1})^2 \left(\frac{1}{N(u_i - u_{i-1})} - 1 \right)^2}{\sum_{i=1}^N (u_i - u_{i-1})^2}}, \end{aligned} \quad (\text{D.13})$$

Hence, for each sample size N , we will obtain 10^6 pairs of d_{KS} and d_{DN} .

D.4.1 KS Distance and Sample Size

The d_{KS} value at different deciles exhibits the power law relationship as $d_{KS} \propto A_{KS} N^{-\alpha_{KS}}$ for $N > 50$, shown in Figure. D.1a. By assuming that the proportionality constant A_{KS} depends only on the percentile level $\wp_{KS}$, we begin experimenting

with several functional forms and eventually settled for

$$A_{KS} = \left(\frac{100}{\wp_{KS}} - 1 \right)^{-0.176} \exp(-0.274). \quad (\text{D.14})$$

The value of the exponent α_{KS} can be determined by the slopes of the different deciles. For simplicity, we assume constant exponent,

$$\alpha_{KS} = 0.492, \quad (\text{D.15})$$

which is the mean of the slopes. In Figure. D.1b, we fit the A_{KS} and α_{KS} to the different percentiles $\wp_{KS}$, which provided insight when choosing the functional form for α_{KS} and A_{KS} . Combining Eq. (D.14) with Eq. (D.15), we obtain the d_{KS} and N at percentile $\wp_{KS}$,

$$d_{KS}(\wp_{KS}, N) = \frac{\left(\frac{100}{\wp_{KS}} - 1 \right)^{-0.176} \exp(-0.274)}{N^{0.492}}, \quad N > 50. \quad (\text{D.16})$$

Figure D.1: (a) Log-log plot of d_{KS} against N for different deciles going from the 10th percentile (blue) to the 90th (red), obtained from 10^6 simulations. (b) Fits for A_{KS} and α_{KS} at different percentile.

D.4.1.1 Analytical Solution Relating Distribution Noise to Sample Size

For a sample distributed by $U(0,1)$ we are able to sort it to get $U^{(s)}$, where the superscript (s) refers to the ordered sample. We now define $\Delta = \{\Delta_1 = u_1^{(s)}, \Delta_2 = u_2^{(s)} - u_1^{(s)}, \Delta_3 = u_3^{(s)} - u_2^{(s)}, \dots, \Delta_N = u_N^{(s)} - u_{N-1}^{(s)}\}$. The Δ is the spacing of ordered

uniform distribution, which is $f_{\Delta}(\Delta) = N(1 - \Delta)^{N-1}$ distributed [153]. Rewriting Eq. (D.13) in terms of Δ yield

$$d_{DN} = \sqrt{\frac{\sum_{i=1}^N \left(\Delta_i \left(\frac{1}{N\Delta_i} - 1 \right) \right)^2}{\sum_{i=1}^N \Delta_i^2}}, \quad (\text{D.17})$$

in the limit where N is large ($N > 50$), we can apply the continuous approximation,

$$\begin{aligned} d_{DN}^{(*)} &= \sqrt{\frac{\int_0^1 \left(\Delta \left(\frac{1}{N\Delta} - 1 \right) \right)^2 N(1 - \Delta)^{N-1} d\Delta}{\int_0^1 \Delta^2 N(1 - \Delta)^{N-1} d\Delta}} \\ &= \sqrt{\frac{1}{2} + \frac{2 - N}{2N^2}}. \end{aligned} \quad (\text{D.18})$$

where the $(*)$ indicates the analytically derived distribution noise, we have performed the integration in the surd, using the properties of the Beta distribution. Eq. (D.18) shows $d_{DN}^{(*)} \rightarrow 1/\sqrt{2}$ as $N \rightarrow \infty$ as illustrated in Figure. D.2.

Figure D.2: Relationship between distribution noise d_{DN} and sample size N at deciles going from the 10th percentile (blue) to the 90th (red), obtained from 10⁶ simulations. The d_{DN} value converges to $1/\sqrt{2}$ as N increases.

D.4.2 Distribution Noise and Sample Size

We show the absolute deviation of d_{DN} value for the expected value $d_{DN}^{(*)}$ at different deciles Figure. D.3. Again there are straight lines on log-log plot, which indicate the power law relationship as $d_{DN} - \langle d_{DN} \rangle \propto A_{DN} N^{-\alpha_{DN}}$ for $N > 50$.

Figure D.3: Relationship between absolute deviation of distribution noise from expected value, $|(d_{DN} - d_{DN}^{*})|$ and sample size N at various deciles obtained from 10^6 simulations. The top figure shows deciles from the 10^{th} percentile (blue) to to the 40^{th} (green), while the bottom shows decile from the 60^{th} (green) to the 90^{th} (red).

Figure D.4: Fits for A_{DN} and α_{DN} at different percentile.

Thus, by assuming the proportionality constant A_{DN} depends only the percentile level $\wp_{DN}$, after experimenting with several functional forms, we write down the relationship between d_{DN} and N at percentile $\wp_{DN}$ as

$$A_{DN} = \Phi(\wp_{DN} - 50) \exp \left(-\frac{[50 - |\wp_{DN} - 50|]^{0.430}}{|\wp_{DN} - 50|^{0.302}} \right), \quad (\text{D.19})$$

where $\Phi(x)$ represents the sign of x . For simplicity we assume constant power law exponent $\alpha_{DN} = 0.495$, which is average of the slopes in log-log plot excluding $\wp_{DN} \in [40, 60]$. Refer to Figure. D.4 for the fits of proportionality constant and power law exponent at different percentile. Hence, the equation that relate d_{DN} and N at percentile $\wp_{DN}$ is

$$d_{DN}(\wp_{DN}, N) = \langle d_{DN} \rangle + \Phi(\wp_{DN} - 50) \frac{\exp \left(-\frac{[50 - |\wp_{DN} - 50|]^{0.430}}{|\wp_{DN} - 50|^{0.302}} \right)}{N^{0.495}}, \quad (\text{D.20})$$

where $\Phi(x)$ represents the sign of x , and

$$\langle d_{DN} \rangle = \sqrt{\frac{1}{2} + \frac{2 - N}{2N^2}} \left(\frac{N}{N + 0.5} \right), \quad (\text{D.21})$$

is the mean of d_{DN} that converges to $1/\sqrt{2}$ as $N \rightarrow \infty$, and $N/(N + 0.5)$ is added in as a correction factor for small N . Again this result also suggests that if we have two samples with sizes N_1 and N_2 with $N_2 > N_1$ from the same distribution, we should compare $N_1^{0.495}(d_{DN}^{(1)} - 1/\sqrt{2})$ against $N_2^{0.495}(d_{DN}^{(2)} - 1/\sqrt{2})$. Otherwise, we will make wrong conclusion that the N_2 sample fits the distribution better.

D.4.3 Correlation Between KS Distance and Distribution Noise

Since d_{KS} and d_{DN} are both measure of deviations, we need to account for the inherent correlation between these two variables, when applying them simultaneously to significance testing. To do this, we compute the Pearson correlation between d_{KS} and d_{DN} obtained from the 10^6 simulated samples for various sample sizes. Fig. D.5 shows the fits of the correlation coefficients,

$$\rho_{d_{KS}, d_{DN}}(N) = \frac{e}{N^{0.481}}. \quad (\text{D.22})$$

and confirms our intuition that d_{KS} is positively correlated with d_{DN} . Since d_{KS} is a measure at the cumulative level, the correlation between d_{KS} and d_{DN} vanishes as $N \rightarrow \infty$, since the distribution noises cancel each other in this limit.

Figure D.5: Correlation between d_{KS} and d_{DN} at different sample sizes, obtained from the 10^6 simulated samples.

D.5 Fits to Real Data

Figure D.6: Year by year Taiwan income data from 1981 to 1990, fitted to power law distributions. In all figures, the empirical CDF that is adjusted for FLE is shown as black dots, while the unadjusted and adjusted fits are shown as green and red dashed line respectively. $P_{KS/DN}$ -values (in percentage) are for unadjusted (green) and adjusted (red) fits. We separate the p -values obtained using the CSN method (left) from those using inversion formulae (right) by a '/'.

Figure D.7: Year by year Taiwan income data from 1991 to 2001, fitted to power law distributions. In all figures, the empirical CDF that is adjusted for FLE is shown as black dots, while the unadjusted and adjusted fits are shown as green and red dashed line respectively. $P_{KS/DN}$ -values (in percentage) are for unadjusted (green) and adjusted (red) fits. We separate the p -values obtained using the CSN method (left) from those using inversion formulae (right) by a '/'.

Figure D.8: Year by year Taiwan income data from 2002 to 2011, fitted to power law distributions. In all figures, the empirical CDF that is adjusted for FLE is shown as black dots, while the unadjusted and adjusted fits are shown as green and red dashed line respectively. $P_{KS/DN}$ -values (in percentage) are for unadjusted (green) and adjusted (red) fits. We separate the p -values obtained using the CSN method (left) from those using inversion formulae (right) by a '/'.

Figure D.9: Year by year Taiwan income data from 1981 to 1990, fitted to exponential distributions with $x_{max}^{(EXP)} = x_{min}^{(PL)}$. In all figures, the empirical CDF that is adjusted for FLE is shown as black dots, while the unadjusted and adjusted fits are shown as green and red dashed line respectively. $P_{KS/DN}$ -values (in percentage) are for unadjusted (green) and adjusted (red) fits. We separate the p -values obtained using the CSN method (left) from those using inversion formulae (right) by a '/'.

Figure D.10: Year by year Taiwan income data from 1991 to 2001, fitted to exponential distributions with $x_{max}^{(EXP)} = x_{min}^{(PL)}$. In all figures, the empirical CDF that is adjusted for FLE is shown as black dots, while the unadjusted and adjusted fits are shown as green and red dashed line respectively. $P_{KS/DN}$ -values (in percentage) are for unadjusted (green) and adjusted (red) fits. We separate the p -values obtained using the CSN method (left) from those using inversion formulae (right) by a '/'.

Figure D.11: Year by year Taiwan income data from 2002 to 2011, fitted to exponential distributions with $x_{max}^{(EXP)} = x_{min}^{(PL)}$. In all figures, the empirical CDF that is adjusted for FLE is shown as black dots, while the unadjusted and adjusted fits are shown as green and red dashed line respectively. $P_{KS/DN}$ -values (in percentage) are for unadjusted (green) and adjusted (red) fits. We separate the p -values obtained using the CSN method (left) from those using inversion formulae (right) by a '/'.

Figure D.12: Year by year SGX normalized returns from 2009 to 2016, fitted to power law distributions. In all figures, the empirical CDF that is adjusted for FLE is shown as black dots, while the unadjusted and adjusted fits are shown as green and red dashed line respectively. $P_{KS/DN}$ -values (in percentage) are for unadjusted (green) and adjusted (red) fits. We separate the p -values obtained using the CSN method (left) from those using inversion formulae (right) by a '/'.