Communications of the Korean Statistical Society Vol. 15, No. 2, 2008, pp. 173–191

서울시 아파트 가격 추세의 모형화

황은연[1], 권용찬[2], 장동익[2], 이재용[3], 오희석[3]

요 약

본 연구의 목적은 최근의 서울시 아파트 가격 추세를 동적선형모형에 적합하여 분석하는 것이다. 이를 위하여 국민은행에서 제공하는 "KB 아파트 시세" 자료의 서울시 30평대별 아파트 평당 매매가를 이용하였다. 시점은 2003년 6월 24일부터 2006년 8월 28일까지로 제한하였다. 탐색적 자료분석을 통해 서울시 30평대 아파트 가격의 추세는 크게 두 지역으로 구분되어지는 것을 알 수 있었다. 아파트 가격은 구분된 지역의 공통 추세와 각 구별 특성으로 표현될 수 있다고 가정하고, 베이즈적 방법인 MCMC를 이용하여 동적선형모형을 추정하였다.

주요용어: 서울시 아파트; 동적선형모형; MCMC.

1. 서론

최근 몇 년간 부동산 열풍은 대한민국 전체를 흔들어놨다고 해도 과언이 아니다. 어른이나 아이 누구를 막론하고 부동산 가격과 정부 정책에 관심을 가지고 이야기하는 전 국민의 부동산 박사화가 되어버린 것이다. 특히 강남과 신도시를 중심으로 급증하는 아파트 가격에 정부는 일주일이 멀다하고 각종 정책을 발표하고 있지만 그 열기를 식히기에는 역부족으로 보인다.

이처럼 수도권 아파트 가격이라는 단어가 국민 최대의 관심사가 되어왔음에도 불구하고 이에 대한 연구는 특정지역 아파트의 단기간 가격변화 추이를 수치화하여 보여주는데 지나지 않고 있다. 수치로 해석된 증가율은 현 시점의 상태를 인식하기에는 적절할지 모르나, 앞으로의 가격변화 추세를 예측하기에는 적합하지 않다.

이에 본 연구에서는 서울시 아파트 가격 추세를 예측하는데 있어 도움이 되는 추세가 있는지를 알아보고자 한다. 여기에서는 2003년 6월부터 2006년 8월까지 3년 2개월 간의 서울시 30평형대 아파트 가격 자료를 기준으로 하여 아파트 가격 변화의 추세를 분석하였다. 아파트 가격 자료는 서울시 25개 구의 구별 아파트 평당 매매가를 평활하여 이용하였다. 자료는 "KB 아파트 시세 DB"에 나와있는 모든 동을 조사하였으나, 동 단위의 구분이 명확치 않아 지역구분을 구 단위로 하였다.

본격적인 분석에 앞서 탐색적 자료분석을 통해 구별 아파트 가격의 변화 양상을 파악한 결과, 서울시 25개 구는 크게 두 지역으로 구분될 수 있다는 것을 알 수 있었고, 아파트 가

1) (151-747) 서울시 관악구 신림 9동, 서울대학교 통계학과, 석사과정. 교신저자: victorf@paran.com
2) (151-747) 서울시 관악구 신림 9동, 서울대학교 통계학과, 박사과정.
3) (151-747) 서울시 관악구 신림 9동, 서울대학교 통계학과, 교수.

격 추세가 구분된 지역의 공통 특성과 각 구별 특성 모두에 의해 영향을 받는다고 보여졌다. 이에 위의 특성을 만족하는 베이지안적 시계열 모형인 동적선형모형(Dynamic Linear Model, DLM)을 기본 모형으로 하여 지역 구분과 등분산 유무에 따른 4가지 적합 모형에 자료를 적합시켜 보고 그 중 어떤 모형이 자료에 가장 적합한지를 판단해 보고자 한다. 자료의 적합에는 베이지안 데이터 분석을 위한 통계 패키지인 WinBUGS를 이용하였다.

우선 2장에서 자료의 수집과정에 대해 간략히 소개한 다음, 3장에서는 구체적으로 모형에의 적합과 분석의 결과를 보고자 한다. 서울시 25개 구를 어떻게 두 지역으로 구분하였는지를 3.1절에서 보이고, 3.2절에서는 분석에 사용한 기본모형인 동적선형모형에 대해 설명하고자 한다. 다음으로 동적선형모형을 이용한 4가지의 대안 모형을 3.3절에서 각각 정의하고, 앞의 4가지 모형 중 어느 모형이 자료에 가장 적절한지를 3.4절에서 판단하고자 한다. 마지막으로 4장에서는 선택된 모형에서 구해진 추정값을 가지고 시계열 그림을 그려 실제 관측값과 비교하면서 모형적합이 적절한지를 판단해 보고자 한다.

본 연구의 목적은 앞서 언급하였듯이 서울시 아파트 가격 추세를 예측하는데 있어 도움이 되는 추세의 존재유무와 그 적합성을 판단해 보는 것으로, 아파트 가격에 영향을 미치는 구체적인 변수에의 영향에 대한 부분은 언급하지 않고자 한다.

2. 자료 개요

연구에서는 국민은행 홈페이지에 공개된 "KB 아파트 시세 DB"의 아파트 시세자료를 이용하였다. 기간은 2003년 6월 24일부터 2006년 8월 28일까지 3년 2개월 동안의 기간(164주)을 대상으로 하였다.

본 연구에서는 서울시 자료만을 사용하므로 전체 자료 중 서울시 25개 구의 210개 동으로 대상범위를 한정하였고, 평형대는 30평형대로 제한하였다. 표본추출은 아파트 가격 변화율에 따른 층화표본추출을 사용하였으며, 대표성을 위하여 최소한 한 동에서 2개 이상의 아파트 시세를 조사하였다. 아파트 시세는 평형에 따른 차이를 없애기 위해 매매가를 평수로 나눈 평당 매매가를 이용하였고, 분석에 있어서는 이를 평활한 평당 매매가의 구별 평균을 구하여 사용하였다.

그림 2.1은 서울시 25개 구의 30평형대 아파트 시세 조사를 위한 표본추출 지역을 나타낸 것이다.

3. 모형 적합과 분석

3.1. 구별 평균 시계열 그림의 분석

우선 서울시 25개 구의 시간에 따른 구별평균 아파트 시세는 그림 3.1과 같다. 그림을 보면 구별 아파트 시세가 실선과 점선으로 표시되어 있는데, 실선으로 표시된 구별 아파트 시세와 점선으로 표시된 구별 아파트 시세 간에는 뚜렷한 차이가 존재함을 알 수 있다. 이로부터 서울시의 25개구는 실선으로 표시된 뚜렷한 증가추세와 step function 형태를 가지

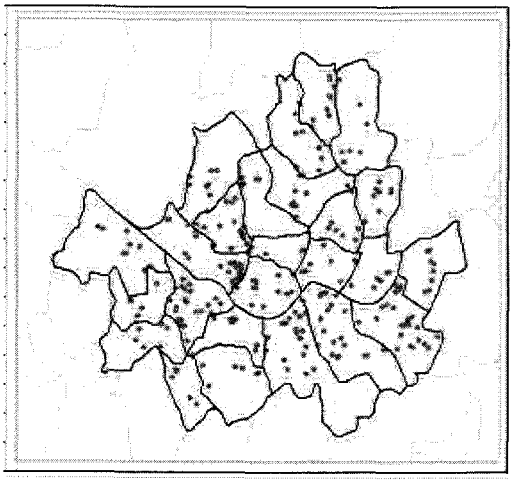

그림 2.1: 표본추출지역

고 있는 지역 1 (실선)과 증가추세를 보이지 않고 비교적 정체되어 있는 지역 2 (점선)의 두 지역으로 구분할 수 있게 된다.

그러므로 서울시 25개 구는 실선으로 표시된 지역 1(광진구, 마포구, 성동구, 성북구, 용산구, 은평구, 중구, 강남구, 강동구, 강서구, 구로구, 동작구, 서초구, 송파구, 양천구, 영등포구의 16개 구)과 점선으로 표시된 지역 2(강북구, 노원구, 도봉구, 동대문구, 종로구, 서대문구, 중랑구, 관악구, 금천구의 9개 구)의 두 지역으로 구분되어진다.

그림 3.1을 그림 3.2와 같이 지역 1과 지역 2로 각각 구분하여 시계열 그림을 그려보면, 지역 1과 지역 2 간 뚜렷한 차이가 존재한다는 것을 다시 한 번 확인할 수 있다. 지역 1은 아파트 시세가 계단 형식의 증가 추세를 보이고 있고, 지역 2는 비교적 정체되어 있다는 것을 알 수 있다. 또한 지역 1과 지역 2는 각각에 있어 공통적인 추세가 존재함을 알 수 있고, 구마다 공통 특성과 함께 각각의 구별 특성에 의해 영향 받는다는 것을 알 수 있다. 이에 각 지역의 공통적인 추세를 가장 잘 나타낸다고 보여지는 강남구(지역 1)와 동대문구(지역 2)를 지역의 추세를 대표하는 구로 선택하여 3장에서 제시할 동적선형모형에서의 공통추세 값으로 사용하였다.

3.2. 동적선형모형과 DIC

(1) 동적선형모형

아파트 가격변화 추세의 적합에 사용된 동적선형모형은 일반적으로 상태들의 시간에 따른 동적 변화를 나타내는 구조 방정식과 상태값에 의해 조건부된 관측값의 표본 분포를 나타내는 관측 방정식, 그리고 초기 정보들로 이루어져 있다. 고정된 모수를 가정하는 일반적인 시계열 모형에 비해, 동적선형모형은 시간에 따라 변화하는 모수를 가정하고 있다.

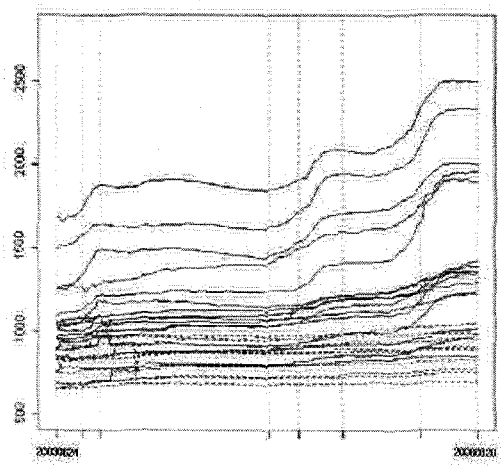

그림 3.1: 서울시 25개구 구별 평균 아파트 시세 시계열 그림 (가로축은 시간(주 단위)을 나타내고, 세로축은 구별 평균 아파트 시세(단위: 만원)를 나타낸다.)

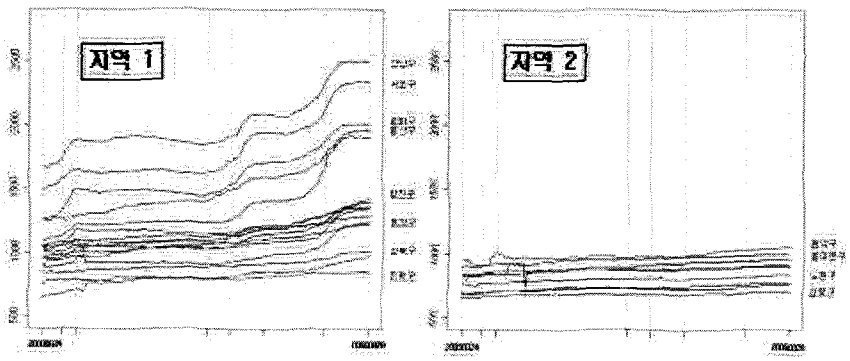

그림 3.2: 지역 1과 지역 2로 구분하여 그린 구별 평균 아파트 시세 시계열 그림

그리고 구조 모형과 관측 모형은 각 시점에서 선형의 성질을 만족하고 있지만, 모수들은 바로 전 시점의 정보에 의해 추정되고 또한 구조 모형이 관측 모형에 영향을 미치므로, 동적선형모형은 결과적으로 비선형적 특징을 보이게 된다.

또한 대부분의 상태공간적 모형이 최대가능도 추정법에 의존하는 것과는 달리, 동적선형모형은 각 시점에서의 모수 추정에 베이지안 추정 방법을 사용하고 있으므로 베이지안적 시계열 모형이라고 할 수 있다 (Bernado와 Smith, 1994).

동적선형모형은 다음과 같이 t시점에서의 관측값 y_t를 나타내는 관측 방정식과 t시점에서의 동적계수인 상태벡터 μ_t를 나타내는 구조방정식, 그리고 초기 정보로 이루어져 있

다 (West와 Harrison, 1997).

$$\begin{aligned} \text{관측 방정식:} \quad & y_t = \eta_t \mu_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_\epsilon^2), \quad t = 1, 2, \ldots, N, \\ \text{구조 방정식:} \quad & \mu_t = M_t \mu_{t-1} + \nu_t, \quad \nu_t \sim N(0, \sigma_\nu^2), \quad t = 1, 2, \ldots, N, \\ \text{초기 정보:} \quad & (\mu_0 \mid D_0) \sim N(m_0, C_0). \end{aligned}$$

관측 방정식은 관측값과 설계벡터 사이의 관계에 대해 정의하고 있다. y_t는 t시점에서의 관측값이고, η_t는 t시점에서의 $p \times 1$인 설계벡터이다. y_t는 μ_t가 주어졌을 때 다른 모든 관측값들과 모수값들에 대해 독립이고, 이는 일반적으로 현재 상태에서 미래는 과거와 독립임을 의미한다. ϵ_t는 평균이 0이고 분산이 σ_ϵ^2인 정규분포를 따르는 관측오차이다.

구조 방정식은 시간에 따라 변화하는 모수에 대해 설명하고 있다. 여기에서 μ_t는 t시점에서의 $p \times 1$인 상태벡터이고, M_t는 t시점에서 상태벡터의 변화를 설명하는 전이행렬이다. μ_{t-1}이 주어졌을 때, μ_t의 분포는 y_{t-1}의 값과는 독립적으로 결정되며, M_t는 보통 알려져 있거나 시간에 따라 변하지 않는 고정된 값을 가지고 있다. ν_t는 평균이 0이고 분산이 σ_ν^2인 정규분포를 따르는 구조오차이며, 관측 방정식에서의 ϵ_t와 ν_t는 상호독립이다.

초기 정보는 시점 $t = 0$에서 μ_0의 수준에 대한 예측자 확신에의 확률적 표현이다. m_0는 사전 상태벡터의 평균벡터를 나타내고, C_0는 분산-공분산 행렬을 나타낸다. D_0는 사전 관측값과 모수의 분포정보 등 모든 사전정보들로 구성된 집합이다.

이와 같이 동적선형모형을 이루는 두 방정식은 각각에 있어서는 선형이지만 구조 방정식이 관측 방정식에 영향을 주므로 비선형의 특징을 보이게 되므로, 선형과 비선형 모형의 성질을 모두 가지고 있다고 할 수 있다.

(2) DIC(Deviance Information Criterion)

3.3절에서 동적선형모형을 이용한 4가지의 대안 모형 중 어느 모형이 자료에 가장 적합한지를 판단함에 있어 본 연구에서는 DIC를 기준으로 채택하였다.

DIC는 베이지안 통계학에서 모형 비교에 이용되는 대표적 통계량으로서 자료의 모형에 대한 적합도와 모형의 복잡도를 모두 고려하는 통계량이다. Spiegelhalter 등 (2002)에 의해 제안된 통계량으로 DIC는 다음의 원리에 기초하고 있다.

$$\text{DIC} = \text{적합도(goodness of fit)} + \text{복잡도(complexity)}$$

편차를 통한 적합도와 유효모수 추정값에 의한 복잡도는 다음과 같이 측정한다.

$$\begin{aligned} \text{적합도:} \quad & D_\theta = -2 \log L(data \mid \theta), \\ \text{복잡도:} \quad & p_D = E_{\theta|y}[D] - D(E_{\theta|y}[\theta]) \; = \bar{D} - D(\bar{\theta}). \end{aligned}$$

따라서 DIC는 다음과 같이 AIC와 유사하게 정의될 수 있다.

$$\text{DIC} = D(\bar{\theta}) + 2p_D \; = \; \bar{D} + p_D.$$

DIC는 그 값이 작을수록 모형이 자료를 잘 설명한다고 할 수 있다. 즉 작은 DIC 값을 갖는 모형일수록 자료에 적합하다고 판단할 수 있다. 또한 DIC는 본 연구에서 사용하는 베이지안 데이터분석 소프트웨어인 WinBUGS의 MCMC 실행에 의해 쉽게 계산될 수 있고, WinBUGS의 Inference/DIC 메뉴를 이용하여 계산된 값을 확인할 수 있다.

3.3. 추정 모형

앞에서 언급하였듯이 서울시 25개 구의 아파트 시세는 탐색적 자료분석에 의해 지역 1과 지역 2로 나누어질 수 있고, 각 지역에 있어서는 시간에 따라 변하는 공적변화 추세가 존재함을 알 수 있었다. 이에 동적선형모형을 기본 모형으로 적합하고자 한다. 모형에서는 각 지역의 공통적인 추세를 가장 잘 나타낸다고 보여지는 강남구(지역 1)와 동대문구(지역 2)를 지역의 추세를 대표하는 구로 선택하여 동적선형모형에서의 공통추세 값으로 사용하였다.

동적선형모형을 기본으로 하는 적합 모형은 지역 구분과 등분산의 유무에 따라 다음의 4가지 모형으로 정의해 볼 수 있다. 모형 1은 공통적인 동적변화추세가 지역마다 존재하고 분산도 지역마다 다르다고 가정한 모형이고, 모형 2는 공통적인 동적변화추세는 지역마다 존재하되 분산은 지역에 구분없이 같다고 가정한 모형이다. 모형 3은 지역을 따로 구분하지 않은 상태에서 공통적 동적변화추세를 강남구의 시세로 가정한 모형이고, 모형 4는 지역을 구분하지 않은 상태에서 공통적 동적변화추세를 동대문구의 시세로 가정한 모형이다. 이러한 4가지 모형을 자료에 적합시켜 보고 그 결과를 바탕으로 서울시 아파트 가격 자료의 적합에는 어떤 모형이 가장 적절한지 알아보도록 하겠다. 모형을 정의할 때 강남구를 기준으로 적합시킨 지역 1은 $k = 1$으로 표시하고, 지역 2는 동대문구를 기준으로 적합시킨 지역으로 $k = 2$로 표시하도록 하겠다.

(1) 모형 1: 공통적인 동적변화추세가 지역마다 존재하고 분산도 지역마다 다르다고 가정한 모형

모형 1은 서울시가 지역 1과 2로 구분되고 나누어진 지역은 각각의 공통 추세를 가지며 서로 다른 분산을 가진다고 가정한 모형이다. 각 지역에서 기준이 되는 구는 위에서 언급하였듯이 강남구(지역 1)와 동대문구(지역 2)로 하였고, 이들 구의 자료를 동적선형모형에서의 공통추세의 값으로 사용하였다. 모형적합에는 WinBUGS 프로그램을 사용하였다. 자료는 지역 1과 지역 2의 순서로 입력하되, 기준이 되는 구인 강남구와 동대문구의 자료를 각 지역별 자료의 제일 앞에 놓고 적합하였다. 식으로 표현하면 다음과 같다.

$$\begin{aligned} y_{k,i,t} &= \eta_{k,i}\mu_{k,t} + \epsilon_{k,i,t}, \quad \epsilon_{k,i,t} \sim N(0, \sigma^2_{\epsilon_k}), \\ \mu_{k,t} &= \mu_{k,t-1} + \nu_{k,t}, \quad \nu_{k,t} \sim N(0, \sigma^2_{\nu_k}), \\ \eta_{k,1} &= 1, \quad \sigma^2_{\epsilon_1} \neq \sigma^2_{\epsilon_2}, \end{aligned}$$

$$\eta_{k,i} \sim N(0, 0.1), \quad \tau_{y_k} \sim \Gamma(0.001, 1000), \quad \tau_{\mu_k} \sim \Gamma(0.001, 1000).$$

위에서 k는 지역을 나타내므로 $1,2$의 값을 갖고, i는 지역 $1(i = 16)$과 지역 $2(i = 9)$에 속하는 구를 나타낸다. t는 시간을 나타내는데, 주 단위로 자료가 주어졌으므로 1부터 164까지의 값(164주)을 갖는다. $y_{k,i,t}$는 지역 k, i번째 구의 시점 t에서의 관측값이고, $\mu_{k,t}$는 지역 k의 t시점에서의 상태벡터이다. 공통추세의 값으로 사용되는 강남구의 동대문구의 자료는 각 지역 자료의 처음에 위치하므로 $\eta_{k,1} = 1$이 된다($\eta_{k,1} = 1,\ k = 1,2$). 분산은 지역마다 다르므로 $\sigma^2_{\epsilon_1} \neq \sigma^2_{\epsilon_2}$이 된다. 그리고 $\eta_{k,i}, \tau_{y_k}, \tau_{\mu_k}$는 초기정보로서 모수들의 확률분포를 나타낸다. $\eta_{k,i}$는 평균이 0이고 분산이 0.1인 정규분포를, 관측값의 사전 정밀도를 나타내는 τ_{y_k}는 $\alpha = 0.001$, $\beta = 0.001$을 가지는 감마분포를, 상태벡터의 사전 정밀도를 나타내는 τ_{μ_k}는 $\alpha = 0.001$, $\beta = 0.001$을 가지는 감마분포를 따른다고 가정하였다.

(2) 모형 2: 공통적인 동적변화추세는 지역마다 존재하되 분산은 지역에 구분 없이 같다고 가정한 모형

모형 2는 서울시가 지역 1과 2로 구분되고 지역은 각각의 공통 추세를 가지되 지역 간 동일한 분산을 가진다고 가정한 모형이다. 기준이 되는 강남구와 동대문구 자료는 지역별 자료의 제일 앞에 놓고 적합하였다. 식으로 표현하면 다음과 같다.

$$\begin{aligned} y_{k,i,t} &= \eta_{k,i}\mu_{k,t} + \epsilon_{k,i,t}, \quad \epsilon_{k,i,t} \sim N(0, \sigma^2_\epsilon), \\ \mu_{k,t} &= \mu_{k,t-1} + \nu_{k,t}, \quad \nu_{k,t} \sim N(0, \sigma^2_\nu), \\ \eta_{k,1} &= 1, \quad \sigma^2_{\epsilon_1} = \sigma^2_{\epsilon_2} = \sigma^2_\epsilon, \\ \eta_{k,i} \sim N(0, 0.1), &\quad \tau_y \sim \Gamma(0.001, 1000), \quad \tau_\mu \sim \Gamma(0.001, 1000). \end{aligned}$$

위에서 k는 지역을 나타내므로 $1,2$의 값을 갖고, i는 지역1 $(i = 16)$과 지역 2 $(i = 9)$에 속하는 구를 나타낸다. t는 시간을 나타내는데, 주 단위로 자료가 주어졌으므로 1부터 164까지의 값(164주)을 갖는다. $y_{k,i,t}$는 지역 k, i번째 구의 시점 t에서의 관측값이고, $\mu_{k,t}$는 지역 k의 t시점에서의 상태벡터이다. 공통추세의 값으로 사용되는 강남구의 동대문구의 자료는 각 지역 자료의 처음에 위치하므로 $\eta_{k,1} = 1$이 된다($\eta_{k,1} = 1,\ k = 1,2$). 분산은 지역에 구분없이 같다고 가정하고 있으므로 $\sigma^2_{\epsilon_1} = \sigma^2_{\epsilon_2} = \sigma^2_\epsilon$이 된다. 그리고 $\eta_{k,i}, \tau_y, \tau_\mu$는 초기정보로서 모수들의 확률분포를 나타낸다. $\eta_{k,i}$는 평균이 0이고 분산이 0.1인 정규분포를, 관측값의 사전 정밀도를 나타내는 τ_y는 $\alpha = 0.001$, $\beta = 0.001$을 가지는 감마분포를, 상태벡터의 사전 정밀도를 나타내는 τ_μ는 $\alpha = 0.001$, $\beta = 0.001$을 가지는 감마분포를 따른다고 가정하였다.

(3) 모형 3: 지역을 구분하지 않은 상태에서 공통적 동적변화추세를 강남구의 시세로 가정한 모형

모형 3은 서울시를 두 지역으로 구분하지 않고 하나의 지역으로 보면서 공통 추세는 지역 1의 기준구인 강남구를 따른다고 가정한 모형이다. 기준이 되는 강남구 자료를 제일 앞

에 놓고 적합하였다. 식으로 표현하면 다음과 같다.

$$\begin{aligned} y_{i,t} &= \eta_i \mu_t + \epsilon_{i,t}, \quad \epsilon_{i,t} \sim N(0, \sigma^2_{\epsilon_1}), \\ \mu_t &= \mu_{t-1} + \nu_t, \quad \nu_t \sim N(0, \sigma^2_{\nu_1}), \\ \eta_1 &= 1, \\ \eta_i \sim N(0, 0.1), &\quad \tau_y \sim \Gamma(0.001, 1000), \quad \tau_\mu \sim \Gamma(0.001, 1000). \end{aligned}$$

위에서 i는 지역의 구분 없이 서울시에 속하는 25개 구를 나타낸다. t는 시간을 나타내는데, 주 단위로 자료가 주어졌으므로 1부터 164까지의 값(164주)을 갖는다. $y_{i,t}$는 i번째 구의 시점 t에서의 관측값이고, μ_t는 t시점에서의 상태벡터이다. 공통추세의 값으로 사용되는 강남구의 자료는 가장 앞에 위치하므로 $\eta_1 = 1$이 된다. 그리고 η_i, τ_y, τ_μ는 초기정보로서 모수들의 확률분포를 나타낸다. η_i는 평균이 0이고 분산이 0.1인 정규분포를, 관측값의 사전 정밀도를 나타내는 τ_y는 $\alpha = 0.001$, $\beta = 0.001$을 가지는 감마분포를, 상태벡터의 사전 정밀도를 나타내는 τ_μ는 $\alpha = 0.001$, $\beta = 0.001$을 가지는 감마분포를 따른다고 가정하였다.

(4) **모형 4: 지역을 구분하지 않은 상태에서 공통적 동적변화추세를 동대문구의 시세로 가정한 모형**

모형 4는 서울시를 두 지역으로 구분하지 않고 하나의 지역으로 보면서 공통 추세는 지역 2의 동대문구를 따른다고 가정한 모형이다. 기준이 되는 동대문구 자료를 제일 앞에 놓고 적합하였다. 식으로 표현하면 다음과 같다.

$$\begin{aligned} y_{i,t} &= \eta_i \mu_t + \epsilon_{i,t}, \quad \epsilon_{i,t} \sim N(0, \sigma^2_{\epsilon_2}), \\ \mu_t &= \mu_{t-1} + \nu_t, \quad \nu_t \sim N(0, \sigma^2_{\nu_2}), \\ \eta_1 &= 1, \\ \eta_i \sim N(0, 0.1), &\quad \tau_y \sim \Gamma(0.001, 1000), \quad \tau_\mu \sim \Gamma(0.001, 1000). \end{aligned}$$

위에서 i는 지역의 구분 없이 서울시에 속하는 25개 구를 나타낸다. t는 시간을 나타내는데, 주 단위로 자료가 주어졌으므로 1부터 164까지의 값(164주)을 갖는다. $y_{i,t}$는 i번째 구의 시점 t에서의 관측값이고, μ_t는 t시점에서의 상태벡터이다. 공통추세의 값으로 사용되는 동대문구의 자료는 가장 앞에 위치하므로 $\eta_1 = 1$이 된다. 그리고 η_i, τ_y, τ_μ는 초기정보로서 모수들의 확률분포를 나타낸다. η_i는 평균이 0이고 분산이 0.1인 정규분포를, 관측값의 사전 정밀도를 나타내는 τ_y는 $\alpha = 0.001$, $\beta = 0.001$을 가지는 감마분포를, 상태벡터의 사전 정밀도를 나타내는 τ_μ는 $\alpha = 0.001$, $\beta = 0.001$을 가지는 감마분포를 따른다고 가정하였다.

3.4. 모형 비교

3.4절에서는 3.3절에서 제안한 4가지 모형의 비교를 통해 서울시 아파트 가격 추세에 가장 적합한 모형은 무엇인지 판단해 보고자 한다. 비교의 기준으로는 3.2절에서 제안한

표 3.1: WinBUGS를 통해 구해진 4가지 모형의 DIC값

	모형 1	모형 2	모형 3	모형 4
DIC	39020.0	41670.0	44180.0	44160.0

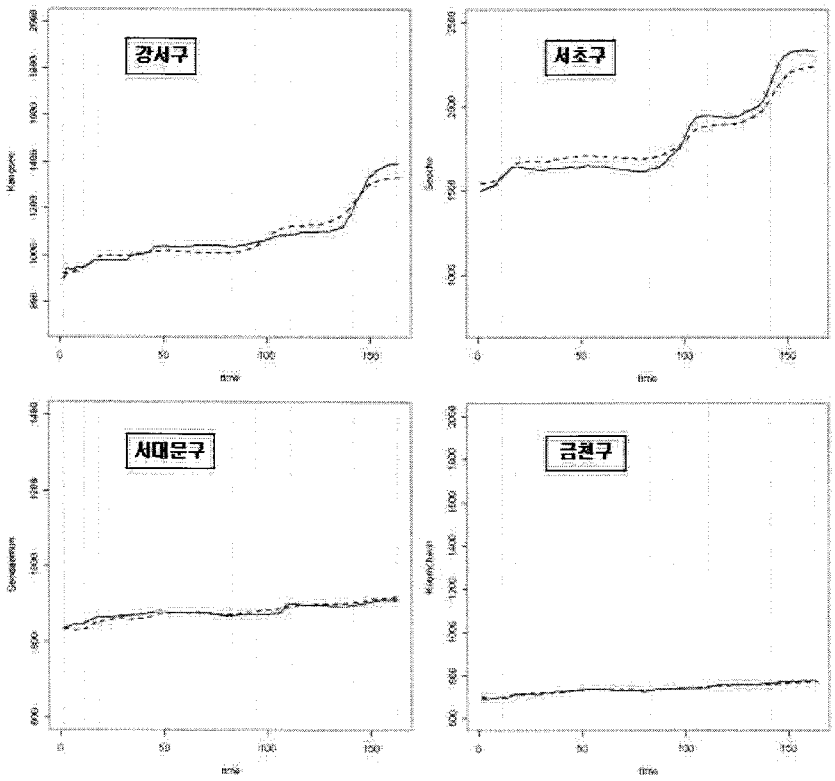

그림 3.3: 관측값과 모형 1에 의한 구별 추정값의 시계열 그림 (가로축은 시간(주 단위)을 나타내고, 세로축은 구별 평균 아파트 시세(단위: 만원)를 나타낸다. 시계열 그림에서 실선은 관측값을 나타내고, 점선은 추정값을 나타낸다.)

베이지안적 모형비교 기준값인 DIC와 추정값에 의한 시계열 그림을 이용하고자 한다. 자료의 적합에는 베이지안 데이터 분석을 위한 통계 패키지인 WinBUGS를 이용하였다. 먼저 WinBUGS를 통해 구해진 4가지 모형의 DIC 값을 알아보면 표 3.1과 같다.

DIC는 그 값이 작을수록 선택된 모형이 자료를 잘 설명하고 있음을 나타낸다. 그러므로 표 3.1로부터 서울시 아파트 가격 자료는 가장 작은 DIC값을 갖는 모형 1에 의해 가장

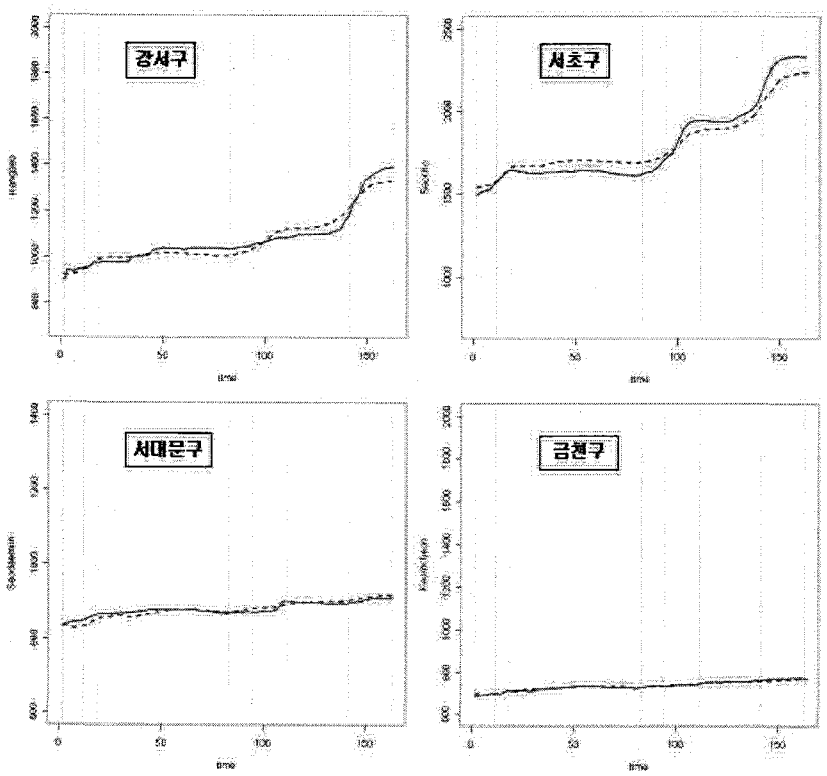

그림 3.4: 관측값과 모형 2에 의한 구별 추정값의 시계열 그림 (가로축은 시간(주 단위)을 나타내고, 세로축은 구별 평균 아파트 시세(단위: 만원)를 나타낸다. 시계열 그림에서 실선은 관측값을 나타내고, 점선은 추정값을 나타낸다.)

잘 설명되고 있다고 말할 수 있다.

다음으로 추정값에 의한 시계열 그림을 살펴보자. 추정값과 관측값을 동시에 표시한 시계열 그림을 통해서도 DIC에서와 비슷한 결론을 얻을 수 있다. 지역 1과 지역 2의 대표적인 4개의 구(강서구와 서초구(지역 1), 서대문구와 금천구(지역 2))에 대하여 앞의 4가지 모형을 적합시켜 얻은 추정값과 관측값의 시계열 그림을 살펴보면 다음과 같다.

그림 3.3은 4개 구의 관측값과 모형 1에 의한 추정값의 시계열 그림이다. 여기에서 강서구와 서초구의 관측값과 모형에 의한 추정값이 동일하지는 않지만 그 추세에 있어서는 유사성을 갖고 있으므로, 구의 전반적인 추세는 동적선형모형에 의해 예측 가능함을 알 수 있다. 서대문구와 금천구의 시계열 그림에서도 추정값과 관측값이 거의 비슷한 값을 가지는 것을 볼 수 있다.

그림 3.4는 4개 구의 관측값과 모형 2에 의한 추정값의 시계열 그림이다. 그림을 보면, 강서구와 서초구는 그림 3.3에서와 같이 전반적 추세는 예측 가능하나 자세한 구별 특성에

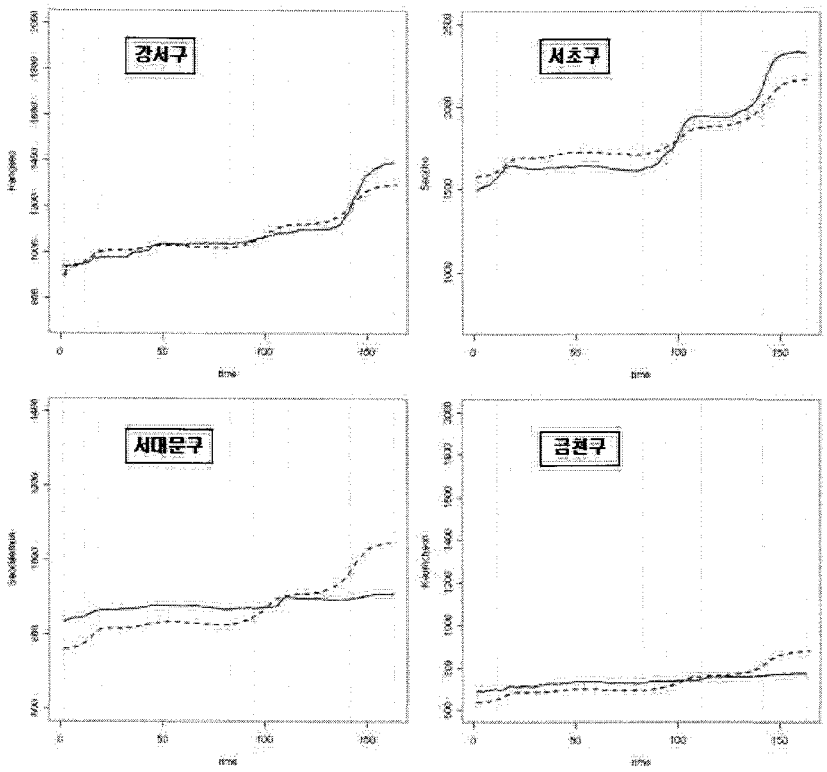

그림 3.5: 관측값과 모형 3에 의한 구별 추정값의 시계열 그림 (가로축은 시간(주 단위)을 나타내고, 세로축은 구별 평균 아파트 시세(단위: 만원)를 나타낸다. 시계열 그림에서 실선은 관측값을 나타내고, 점선은 추정값을 나타낸다.)

대해서는 설명이 안되는 곳이 존재하고 있음을 알 수 있다. 또한 서대문구와 금천구의 시계열 그림에서도 그림 3.3과 같이 추정값과 관측값이 거의 비슷한 값을 가지고 있음을 알 수 있다.

그림 3.5는 4개 구의 관측값과 모형 3에 의한 추정값의 시계열 그림이다. 그림 3.5를 보면, 강서구와 서초구의 관측값과 추정값 간의 차이가 앞의 모형 1·2보다 크다는 것을 알 수 있다. 그리고 서대문구와 금천구에 있어서도 예측력이 모형 1·2에 비해 확연히 떨어짐을 알 수 있다. 이는 지역 2의 추세가 지역 1과 확연히 다름에도 불구하고 지역을 구분하지 않고 지역 1의 공통 추세와 분산을 지역 2에 그대로 적용했기 때문인 것으로 판단된다.

그림 3.6은 4개 구의 관측값과 모형 4에 의한 추정값의 시계열 그림이다. 그림을 보면, 강서구와 서초구의 관측값과 추정값 간의 차이가 그림 3.5와 같이 모형 1·2에 비해 크다는 것을 알 수 있다. 서대문구와 금천구에 있어서도 예측력이 모형 1·2에 비해 확연히 떨어지는 것을 알 수 있다. 이는 지역 1과 지역 2의 추세가 확연히 다름에도 불구하고 지역이 서

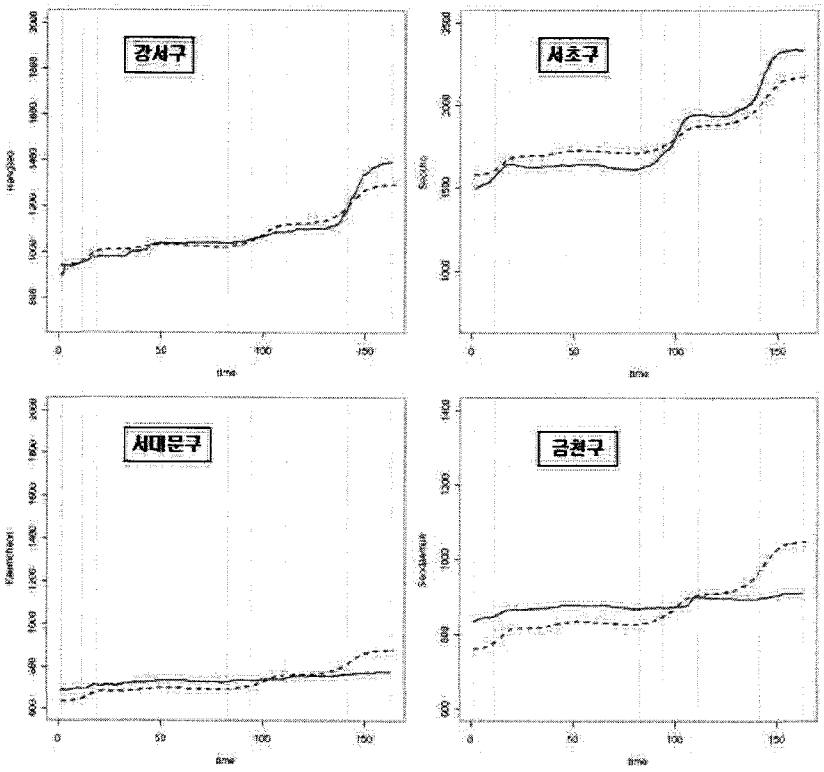

그림 3.6: 관측값과 모형 4에 의한 구별 추정값의 시계열 그림 (가로축은 시간(주 단위)을 나타내고, 세로축은 구별 평균 아파트 시세(단위: 만원)를 나타낸다. 시계열 그림에서 실선은 관측값을 나타내고, 점선은 추정값을 나타낸다.)

로 구분되지 않는다고 가정하여 지역 2의 공통 추세와 분산을 지역 1에 그대로 적용했기 때문인 것으로 판단된다.

지금까지의 시계열 그림에 의한 모형 비교를 통해 서울시 아파트 가격 자료의 적합에는 모형 1과 모형 2가 적절하다고 판단된다. 모형비교의 기준값인 DIC를 이용한 비교에서는 모형 1의 DIC가 모형 2보다 작으므로, 모형 1이 자료에 더 적절하다고 판단된다. 그러므로 서울시 아파트 가격의 추세자료에 있어 가장 적합한 모형은 모형 1이라고 결론내릴 수 있다.

4. 추정값을 이용한 시계열 그림

4장에서는 3장에서 선택된 모형 1을 사용한 관측값과 추정값의 시계열 그림을 서울시 25개 구에 대하여 표현해 보았다.

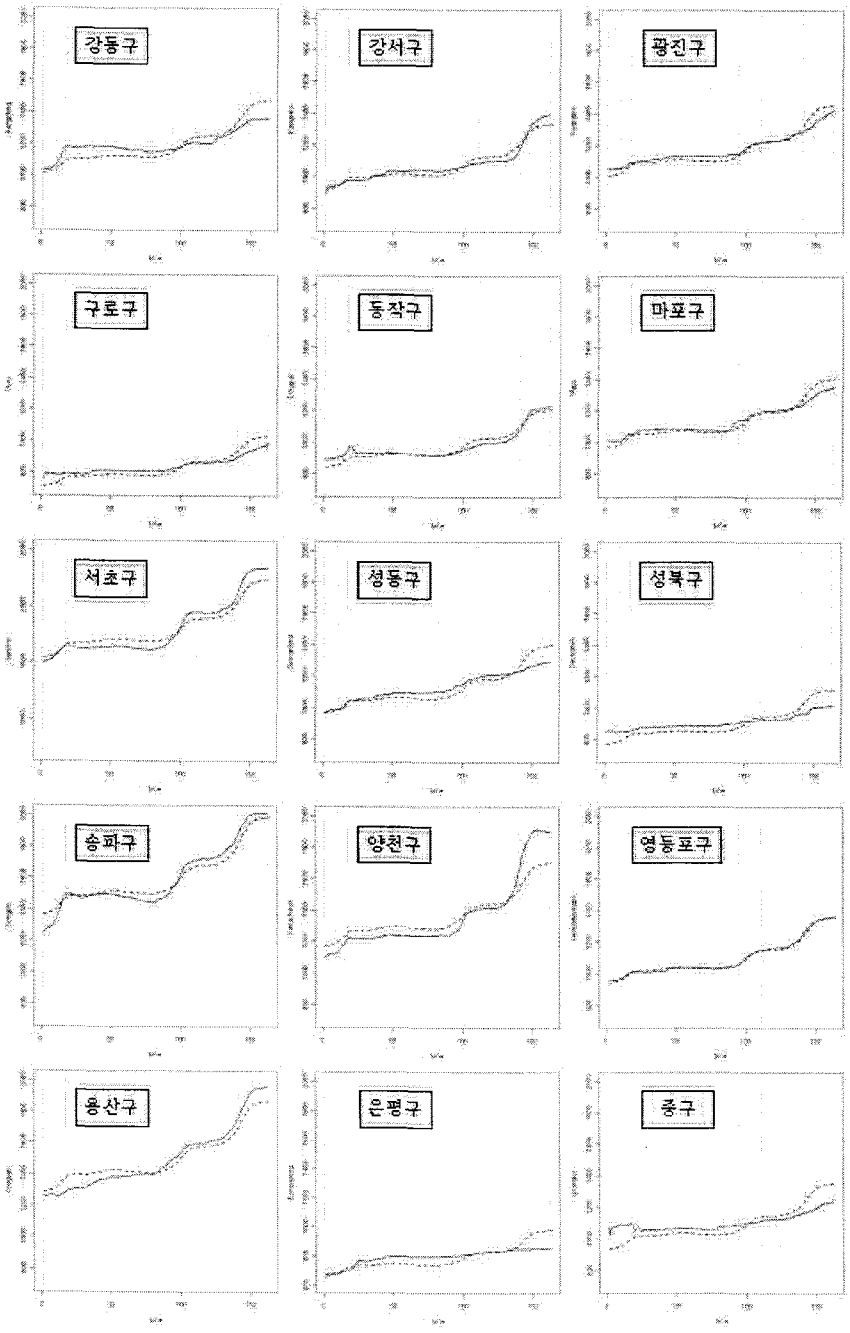

그림 4.1: 지역 1에 속하는 15개 구의 추정값과 관측값의 시계열 그림 (가로축은 시간(주 단위)을 나타내고, 세로축은 구별 평균 아파트 시세(단위: 만원)를 나타낸다. 시계열 그림에서 실선은 관측값을 나타내고, 점선은 추정값을 나타낸다.)

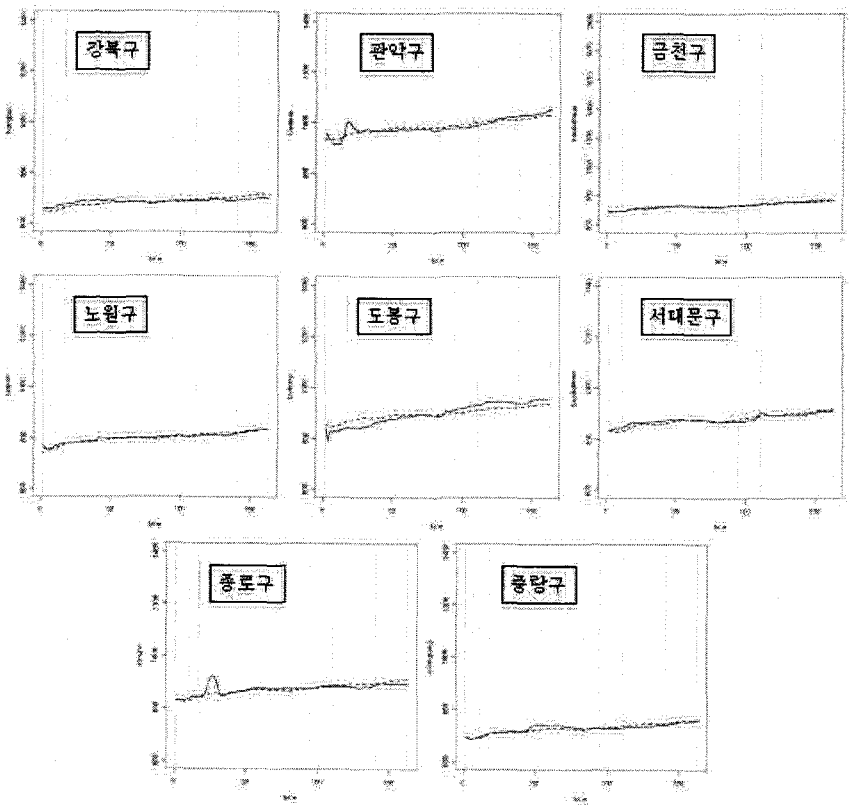

그림 4.2: 지역 2에 속하는 8개 구의 추정값과 관측값의 시계열 그림 (가로축은 시간(주 단위)을 나타내고, 세로축은 구별 평균 아파트 시세(단위: 만원)를 나타낸다. 시계열 그림에서 실선은 관측값을 나타내고, 점선은 추정값을 나타낸다.)

(1) **지역 1 (16개 구)**

지역 1에 속하는 16개구 중 강남구를 제외한 15개구를 모형 1을 이용하여 적합시켰다. 이 때 공통추세 값은 강남구 자료를 사용하였고, 추정값은 WinBUGS를 이용해서 구하였다. 그림 4.1을 보면, 광진구, 마포구, 강서구, 동작구, 영등포구의 5개 구는 모형 1에 의해 잘 적합되고 있음을 알 수 있다. 반면 용산구, 서초구, 양천구의 경우처럼 아파트 가격이 급증하거나, 성북구, 은평구, 중구, 강동구, 구로구와 같이 아파트 가격의 증가폭이 지역 1의 다른 구들에 비해 작은 구들에 대해서는, 전체적인 추세는 모형 1에 의해 설명될 수 있었으나 자세한 변화에 있어서는 적절히 설명될 수 없었음을 알 수 있다.

(2) **지역 2 (9개 구)**

지역 2에 속하는 9개구 중 동대문구를 제외한 8개구를 모형 1을 이용하여 적합시켰다. 이 때 공통추세 값은 동대문구 자료를 사용하였고, 추정값은 WinBUGS를 이용해서 구하

였다. 그림 4.2를 보면, 지역 2는 전반적으로 아파트 가격의 증가폭이 지역 1에 비해 매우 작은 비교적 정체되어 있는 지역이라는 것을 알 수 있다. 또한 증가폭의 구별 차이가 거의 없으므로 대부분의 구가 모형 1에 의해 적합이 잘 되고 있음을 알 수 있다. 그러나 종로구·관악구에서와 같이 초반의 가격이 뛰는 현상은 모형 1에 의해 설명될 수 없었다.

5. 결론 및 고찰

지금까지 서울시 25개 구 30평형대 아파트 가격 추세를 바탕으로 아파트 가격 추세를 예측하는데 도움이 되는 모형이 존재하는지에 대해 살펴보았다.

서울시 25개 구는 탐색적 자료분석을 통해 전반적으로 증가추세를 보이는 지역과 비교적 정체되어 있는 지역의 두 지역으로 구분될 수 있었다. 증가추세가 강하게 보이는 지역 1에는 광진구, 마포구, 성동구, 성북구, 용산구, 은평구, 중구, 강남구, 강동구, 강서구, 구로구, 동작구, 서초구, 송파구, 양천구, 영등포구의 16개 구가 포함되었고, 정체된 추세를 보이는 지역 2에는 강북구, 노원구, 도봉구, 동대문구, 종로구, 서대문구, 중랑구, 관악구, 금천구의 9개 구가 포함되었다.

아파트 가격 자료를 구체적 모형에 적합시켜 보고자 우선 탐색적 자료분석을 하였고, 서울시 아파트 가격은 구분된 지역의 공통 특성과 각각의 구별 특성 모두에 의해 영향 받는다는 것을 알 수 있었다. 따라서 이런 특성을 만족하는 시간에 따라 변하는 모수를 가정하고 있는 동적선형모형이 서울시 아파트 가격 자료의 적합에 적절하다고 판단되었다. 지역 구분과 등분산 유무에 따라 정의된 4가지의 서로 다른 동적선형모형을 이용하여 자료를 적합시켜 본 결과, 공통적인 동적변화추세가 지역마다 존재하고 분산도 지역마다 다르다고 가정한 모형 1이 DIC와 시계열 그림을 모형 비교의 기준으로 하였을 때 자료를 가장 잘 표현하고 있다고 판단할 수 있었다.

선택된 모형 1에 의해 구해진 추정값과 관측값의 시계열 그림을 통해 서울시 구별 아파트 가격의 지역에 따른 전반적인 추세는 설명될 수 있었으나, 가격의 뛰는 현상이나 급등과 같은 자세한 구별 특성은 적합된 모형에 의해 아직 설명될 수 없음을 알 수 있었다.

본 연구의 목적은 앞에서 언급하였듯이 서울시 아파트 가격 추세를 예측하는데 있어 도움이 되는 추세가 존재의 유무와 그 적합성을 판단하는 것이었다. 동적선형모형을 통해 서울시 아파트 가격 추세의 지역에 따른 전반적 추세는 설명할 수 있게 되었으나, 자세한 구별 특성에 있어서는 표현할 수 없는 부분이 아직 존재하고 있다. 이는 향후 아파트 가격 추세에 영향을 미치는 보다 자세한 외생변수를 발견하고 모형에 포함시킴으로써 해결될 수 있을 것으로 판단된다.

부록: Program Code

<모형 1>

```
model {
        tau.x ~ dgamma(0.001, 0.001)
        sigma.x <- (1/tau.x)
        tau.g ~ dgamma(0.001, 0.001)
        sigma.g <- (1/tau.g)
        tau.w ~ dgamma(0.001, 0.001)
        sigma.w <- (1/tau.w)
        tau.d ~ dgamma(0.001, 0.001)
        sigma.d <- (1/tau.d)

        #observation model
         for (i in 2:16){
         for (t in 2:164){
              y[i,t] ~ dnorm(mu.1[i,t], tau.g);
              mu.1[i,t] <- a[i]*x[t];}
              a[i] ~ dnorm(0, 10);}
         for (t in 2:164){
              y[1,t] ~ dnorm(x[t], tau.g);}
         for (i in 18:25){
         for (t in 2:164){
              y[i,t] ~ dnorm(mu.2[i,t], tau.d);
              mu.2[i,t] <- b[i]*w[t];}
              b[i] ~ dnorm(0, 10);}
         for (t in 2:164){
              y[17,t] ~ dnorm(w[t], tau.d);}

        #state model
         for (t in 2:164){x[t] ~ dnorm(x[t-1], tau.x);
                          w[t] ~ dnorm(w[t-1], tau.w);}
         x[1] ~ dnorm(0.0, 0.000001)
         w[1] ~ dnorm(0.0, 0.000001)}
```

<모형 2>

```
model {
        tau.x ~ dgamma(0.001, 0.001)
        sigma.x <- (1/tau.x)
        tau.y ~ dgamma(0.001, 0.001)
```

```
        sigma.y <- (1/tau.y)

        #observation model
         for (i in 2:16){
         for (t in 2:164){
              y[i,t] ~ dnorm(mu.1[i,t], tau.y);
              mu.1[i,t] <- a[i]*x[t];}
              a[i] ~ dnorm(0, 10);}
         for (t in 2:164){
              y[1,t] ~ dnorm(x[t], tau.y);}
         for (i in 18:25){
         for (t in 2:164){
              y[i,t] ~ dnorm(mu.2[i,t], tau.y);
              mu.2[i,t] <- b[i]*w[t];}
              b[i] ~ dnorm(0, 10);}
         for (t in 2:164){
              y[17,t] ~ dnorm(w[t], tau.y);}

        #state model
         for (t in 2:164){x[t] ~ dnorm(x[t-1], tau.x);
                          w[t] ~ dnorm(w[t-1], tau.x); }

         x[1] ~ dnorm(0.0, 0.000001)
         w[1] ~ dnorm(0.0, 0.000001)}
```

<모형 3>

```
model {
        tau.x ~ dgamma(0.001, 0.001)
        sigma.x <- (1/tau.x)
        tau.y ~ dgamma(0.001, 0.001)
        sigma.y <- (1/tau.y)

        #observation model
         for (i in 2:25){
         for (t in 2:164){
              y[i,t] ~ dnorm(mu[i,t], tau.y);
              mu[i,t] <- a[i]*x[t];}
              a[i] ~ dnorm(0, 10);}
         for (t in 2:164){
              y[1,t] ~ dnorm(x[t], tau.y);}

        #state model
        for (t in 2:164){x[t] ~ dnorm(x[t-1], tau.x);}
```

```
        x[1] ~ dnorm(0.0, 0.000001)}
```

<모형 4>

```
model {
        tau.x ~ dgamma(0.001, 0.001)
        sigma.x <- (1/tau.x)
        tau.y ~ dgamma(0.001, 0.001)
        sigma.y <- (1/tau.y)

        #observation model
         for (i in 2:25){
         for (t in 2:164){
              y[i,t] ~ dnorm(mu[i,t], tau.y);
              mu[i,t] <- b[i]*x[t];}
              b[i] ~ dnorm(0, 10);}
         for (t in 2:164){
              y[1,t] ~ dnorm(x[t], tau.y);}

        #state model
        for (t in 2:164){x[t] ~ dnorm(x[t-1], tau.x);}
        x[1] ~ dnorm(0.0, 0.000001)}
```

참고문헌

국민은행 부동산 시세통계, "KB 아파트 시세 DB". http://est.kbstar.com/quics?page=A005877&cc=a042731:a013811.

Congdon, P. (2006). *Bayesian Statistical Modelling.* John Wiley & Sons, New York.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, Ser. B, **64**, 583–639.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models.* Springer-Verlag, New York.

[2007년 6월 접수, 2008년 1월 채택]

Modeling the Trend of Apartment Market Price in Seoul

Eunyeon Hwang[1], Yongchan Kwon[2], Dongik Jang[2], Jaeyong Lee[3], Hee-Seok Oh[3]

Abstract

The goal of this paper is analyzing and modeling the trend of apartment market price in Seoul using the dynamic linear model(DLM). We use the market price per pyeong of 30-pyeong-apartment provided by "KB apartment market price database" of Kookmin bank. The data is collected from June 24^{th}, 2003 to August 28^{th}, 2006. The inspection of the data reveals that the trend of apartment market price in Seoul can be divided into two groups and we assume that the price is expressed by the common trend of divided groups. We try to estimate the price of apartment by DLM using the Bayesian method.

Keywords: Apartment market price; DLM; MCMC.

1) Graduate Student, Department of Statistics, Seoul National University, Seoul 151-747, Korea. Correspondence: victorf@paran.com
2) Graduate Student, Department of Statistics, Seoul National University, Seoul 151-747, Korea.
3) Professor, Department of Statistics, Seoul National University, Seoul 151-747, Korea.