

تحلیل مدل‌های سلسله‌مراتبی بیزی روی داده‌های فضایی- زمانی حجیم قیمت مسکن شهر تهران

امید کریمی* و فاطمه حسینی

دانشگاه سمنان

چکیده: داده‌های قیمت مسکن نسبت به موقعیت قرارگیری آن‌ها در محله‌های مختلف همبسته‌اند و این همبستگی از نوع فضایی (مکانی) است. همچنین قیمت مسکن در ماه‌های مختلف سال متفاوت‌اند، بنا بر این همبستگی از نوع زمانی نیز دارند. برای تحلیل این نوع داده‌ها از مدل‌های فضایی-زمانی استفاده می‌شود. از اهداف مهم در بررسی این نوع از داده‌ها، برازش یک مدل مناسب برای تحلیل فضایی-زمانی آن‌ها، برآورد پارامترها و پیشگویی‌ها در موقعیت‌ها و زمان‌های معلوم است. سه مدل فرایند گاوسی، اتورگرسیو و پویا برای تحلیل داده‌های حجیم قیمت مسکن شهر تهران در نظر گرفته می‌شود. از رهیافت بیزی برای برازش مدل‌ها و پیشگویی در موقعیت‌های جدید و پیش‌بینی در زمان‌های آینده‌ی قیمت مسکن شهر تهران استفاده می‌شود. بهترین مدل با ملاک میانگین مجذور خطای پیشگویی‌ها و مقایسه‌ی زمان محاسبات معرفی می‌شود. در نهایت نقشه‌ی پیشگویی قیمت مسکن تهران در ماه‌های پایان سال ۱۳۹۴ ارائه شده است.

واژگان کلیدی: داده‌های فضایی-زمانی؛ مدل پویا؛ مدل اتورگرسیو؛ مدل فرایند گاوسی؛ رهیافت بیزی.

۱- مقدمه

حجم بالایی از داده‌های فضایی-زمانی به‌طور روزانه و ماهانه جمع‌آوری می‌شوند و مدل‌بندی و تحلیل داده‌های فضایی-زمانی در زمینه‌های علمی گوناگون مورد توجه

* نویسنده‌ی عهده‌دار مکاتبات
دریافت: ۱۳۹۶/۳/۱۲، پذیرش: ۱۳۹۷/۱۱/۱۷.

محققین قرار گرفته است. به عنوان مثال قیمت مسکن شهر تهران به صورت خرید و فروش به طور روزانه در سامانه‌ی بازار املاک ایران ثبت می‌شود. معمولاً برای مدل‌بندی داده‌های فضایی - زمانی از مدل‌های سلسله‌مراتبی بیزی استفاده می‌شود ([۴]، [۶] و [۱۳]). برای تحلیل این مدل‌ها از روش‌های مبتنی بر نمونه‌گیری مانند الگوریتم‌های مونت کارلوی زنجیر مارکوفی^۱ (MCMC) استفاده می‌شود ([۱۴]، [۱۹]).

وجود بعد فضایی و زمانی به طور توأم و اغلب حجم بودن داده‌ها پیچیدگی‌های اساسی بر روی تحلیل و استنباط داده‌ها ایجاد می‌کند به طوری که پیچیدگی محاسبات با الگوریتم‌های MCMC به سرعت افزایش می‌یابد. بنا بر این لزوم مدل‌هایی کارا و سریع برای این گونه از داده‌ها از اهمیت خاصی برخوردار است. کارهای مختلفی در این زمینه انجام شده‌اند که به عنوان مثال می‌توان به کتاب [۸] و منابع معرفی شده توسط آن‌ها اشاره کرد.

مدل‌های فضایی - زمانی پویا به صورت سلسله‌مراتبی توسط [۲۴] معرفی شد. هرچند این مدل‌ها برای تحلیل داده‌های حجمی در سطح وسیعی از ناحیه‌ی مورد مطالعه مناسب نبودند. مشکل اصلی در محاسبه‌ی معکوس ماتریس کوواریانس فضایی و تکرار آن در الگوریتم‌های برازش مدل بود [۱]. کارهای متعددی در این زمینه صورت گرفته است، به عنوان مثال [۷] از کریگیدن محلی^۲ استفاده کرد. یک زیرنمونه‌گیری از بعد فضایی بالا به وسیله‌ی روش پنجره‌ی متحرک توسط [۱۵] پیشنهاد شد و تکنیک‌هایی بر اساس کارایی محاسبات با استفاده از الگوریتم‌های محاسباتی ماتریس پراکنده^۳ معرفی کردند که برای این منظور از تقریب‌های میدان تصادفی مارکوفی گاوسی [۲۰] بهره گرفتند. [۱] یک روش تقریبی ساده با استفاده از فرایند پیشگوی گاوسی^۴ (GPP) برای مدل بیزی داده‌های فضایی چندمتغیره پیشنهاد کردند و روش آن‌ها توسط [۸]، [۹] و [۲۰] بسط داده شد و همچنین [۱۹] مدل‌های اتورگرسیو بیزی را برای داده‌های واقعی ازن به کار گرفتند. [۱۶] به تحلیل بیزی مدل‌های فضایی و فضایی-زمانی برای داده‌های نرخ مرگ و میر به روش تقریب لاپلاس پرداختند و نویسندگان متعددی روی تحلیل بیزی این مدل‌ها مطالعاتی انجام داده‌اند. به عنوان نمونه [۳] و [۱۷] را ببینید.

هدف اصلی در این مقاله مدل‌بندی بیزی داده‌های فضایی-زمانی قیمت مسکن شهر تهران در سال ۱۳۹۴ به صورت ماهانه با استفاده از سه مدل پویا^۵ (DYN)، اتورگرسیو^۶

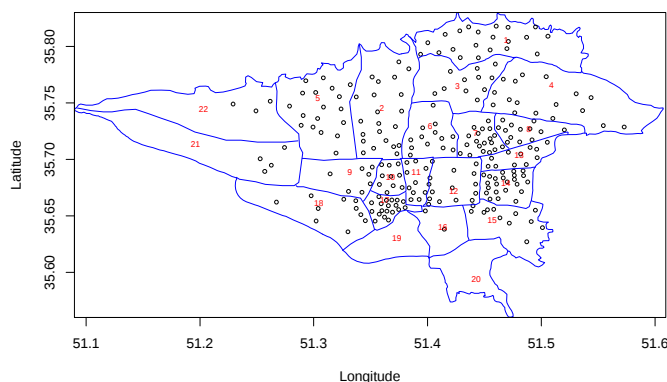
(AR) و فرایند گاوسی^۷ (GP) و در نهایت مقایسه‌ی سه مدل و انتخاب مدل مناسب می‌باشد.

یک تحلیل اکتشافی روی داده‌های قیمت مسکن در بخش ۲ فراهم می‌شود و در بخش ۳ مدل‌های سلسله‌مراتبی بیزی برای داده‌ها معرفی می‌شوند. سپس برازش مدل و پیشگویی فضایی و پیش‌بینی در زمان‌های آینده روی داده‌های قیمت مسکن در بخش ۴ انجام می‌گردد و در انتها بحث و نتیجه‌گیری بیان می‌شود.

۲- تحلیل اکتشافی داده‌های قیمت مسکن

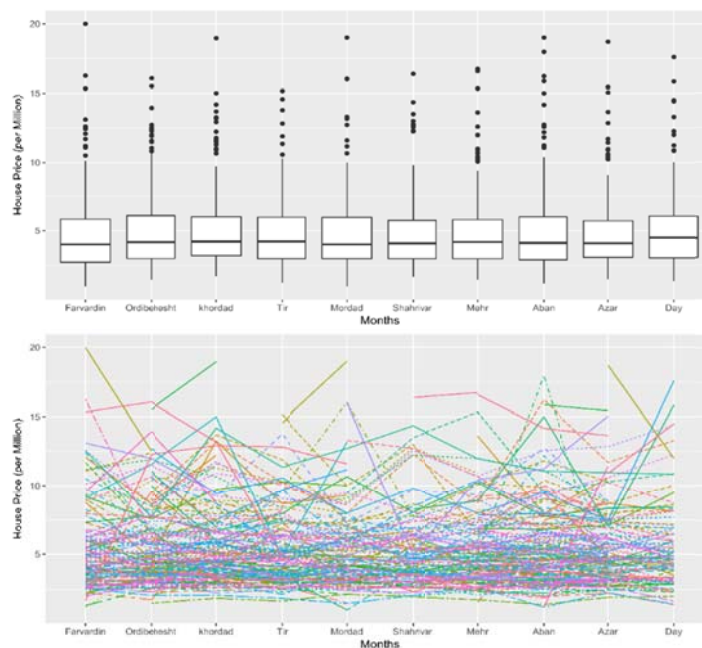
داده‌های قیمت مسکن تهران به صورت روزانه در سامانه‌ی اطلاعات بازار املاک ایران ثبت می‌شود. این داده‌ها را می‌توان از پایگاه www.hmi.mrud.ir مشاهده کرد. قیمت مسکن نوساز (تا یک سال ساخت) برای محله‌های ۲۲ منطقه در ده ماه اول سال ۱۳۹۴ به صورت روزانه استخراج شد.

با توجه به این‌که به صورت روزانه داده‌های گمشده‌ی زیادی برای هر موقعیت (محله‌ها) وجود دارد، متوسط ماهیانه‌ی قیمت مسکن نوساز را برای هر محله‌ی شهر تهران واقع در مناطق ۲۲ گانه محاسبه شد به طوری که ۲۶۰۰ مشاهده (یعنی ۲۶۰ محله در ۱۰ ماه) به دست آمد. موقعیت‌های محله‌ها روی نقشه تهران در شکل ۱ مشخص شده است.



شکل ۱- موقعیت مشاهدات روی نقشه تهران

در شکل ۲ (بالا) نمودار جعبه‌ای قیمت مسکن برحسب ماه‌های سال و شکل ۲ (پایین) نمودار سری زمانی داده‌ها در ۲۶۰ موقعیت رسم شده‌اند. حدود ۳۱ درصد داده‌ها گمشده‌اند و هدف برآورد داده‌های گمشده و پیش‌بینی در ماه‌های بهمن و اسفند در کل ناحیه شهر تهران براساس بهترین مدل فضایی - زمانی است.

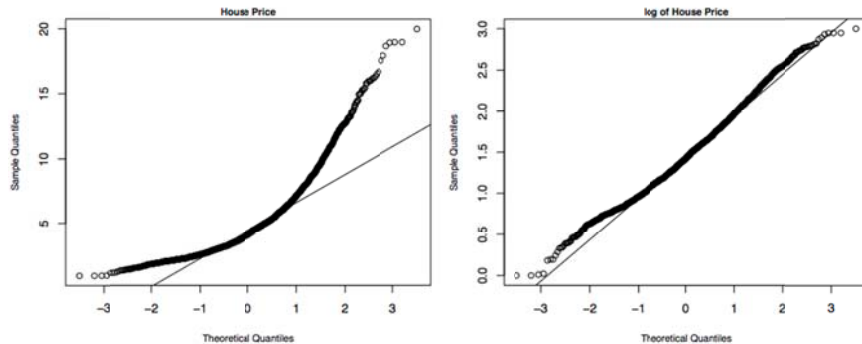


شکل ۲- نمودارهای جعبه‌ای و سری زمانی داده‌های قیمت مسکن برحسب میلیون تومان به متر مربع در ۲۶۰ محله برای ۱۰ ماه از سال ۱۳۹۴

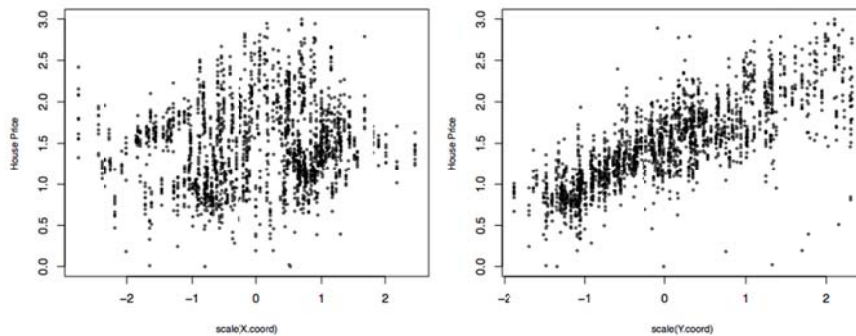
با توجه به نمودار $Q-Q$ نرمال که در شکل ۳ (چپ) مشاهده می‌شود، داده‌های قیمت مسکن چوله به راست هستند و با یک تبدیل لگاریتمی داده‌ها متقارن و به توزیع نرمال نزدیک‌تر می‌شوند (شکل ۳، راست).

نمودار پراکنش برحسب طول و عرض جغرافیایی در شکل ۴ رسم شده‌اند که یک روند خطی با شیب مثبت در جهت جنوب به شمال (عرض جغرافیایی) مشاهده می‌شود.

بنابراین در مدل‌های فضایی - زمانی که در بخش بعد ارایه خواهند شد، از طول و عرض جغرافیایی به عنوان متغیر کمکی استفاده می‌شود.



شکل ۳- (چپ) نمودار Q-Q نرمال قیمت مسکن، (راست) نمودار Q-Q نرمال لگاریتم قیمت مسکن



شکل ۴- (راست) نمودار پراکنش داده‌ها در مقابل طول جغرافیایی و (چپ) نمودار پراکنش داده‌ها در مقابل عرض جغرافیایی

۳- مدل‌ها

فرض کنید $Z_l(s_i, t)$ مقادیر مشاهدات پاسخ در موقعیت s_i ، $i = 1, \dots, n$ و ماه t ام در سال l ام برای $l = 1, \dots, r$ و $t = 1, \dots, T_l$ باشد. در این‌جا دو اندیس زمان به صورت ماه درون سال مشخص شده است که می‌تواند به صورت روز درون سال باشد. قرار دهید.

$$\mathbf{z}_{\ell t} = (z_{\ell}(s_1, t), \dots, z_{\ell}(s_n, t))'$$

و فرض کنید $x_{\ell j}(s, t)$ متغیر کمکی j ام در ماه t ام سال ℓ ام باشد و قرار دهید $X_{\ell t}$ ماتریس $n \times p$ از متغیرهای که شامل عرض از مبدأ است.

[۱۲] مدل‌های فضایی - زمانی بیزی را به صورت سلسله‌مراتبی در سه مرحله بیان کرد:

۱- مشاهدات به شرط یک فرایند تصادفی و پارامترها $(\mathbf{z}|\mathbf{u}, \boldsymbol{\theta})$ ،

۲- فرایند تصادفی به شرط پارامتر $(\mathbf{u}|\boldsymbol{\theta})$ ،

۳- پارامترها $(\boldsymbol{\theta})$.

در ادامه به معرفی سه مدل سلسله‌مراتبی فرایند گاوسی، اتورگرسیو و پویا برای تحلیل داده‌های فضایی زمانی پرداخته می‌شود.

۳-۱- مدل سلسله‌مراتبی فرایند گاوسی

در مرحله‌ی اول مدل به صورت

$$\mathbf{z}_{\ell t} = \mathbf{u}_{\ell t} + \boldsymbol{\varepsilon}_{\ell t}, \quad \ell = 1, \dots, r; t = 1, \dots, T$$

در نظر گرفته می‌شود، که در آن $\boldsymbol{\varepsilon}_{\ell t} = (\varepsilon_{\ell}(s_1, t), \dots, \varepsilon_{\ell}(s_n, t))'$ بردار خطاهای مدل با توزیع $N(\mathbf{0}, \sigma_{\varepsilon}^2 I_n)$ است و σ_{ε}^2 اثر قطعه‌ای یا واریانس خطای تصادفی است، همچنین

$$\mathbf{u}_{\ell t} = (u_{\ell}(s_1, t), \dots, u_{\ell}(s_n, t))'$$

مقادیر واقعی متناظر با $z_{\ell}(s_i, t)$ که شامل مولفه‌های

$$\mathbf{u}_{\ell t} = X_{\ell t} \boldsymbol{\beta} + \boldsymbol{\eta}_{\ell t}$$

است، که در آن $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ بردار پارامترهای ضرایب رگرسیونی،

$\boldsymbol{\eta}_{\ell t} = (\eta_{\ell}(s_1, t), \dots, \eta_{\ell}(s_n, t))'$ اثر تصادفی فضایی - زمانی مدل است که دارای

توزیع $N(\mathbf{0}, \Sigma_{\eta})$ و مستقل از زمان می‌باشد، که در آن $\Sigma_{\eta} = \sigma_{\eta}^2 S_{\eta}$ و σ_{η}^2 پارامتر

واریانس فضایی و S_{η} ماتریس همبستگی فضایی است. در این جا از تابع همبستگی

ماترن ([۱۵]) به صورت

$$K(s_i, s_j; \varphi, \nu) = \frac{1}{\nu^{\nu-1} \Gamma(\nu)} \left(\sqrt{\nu} \|s_i - s_j\| \varphi \right)^\nu K_\nu \left(\sqrt{\nu} \|s_i - s_j\| \varphi \right)$$

استفاده شده است، که در آن K_ν تابع بسل اصلاح‌شده‌ی نوع دوم با مرتبه‌ی ν و $\|s_i - s_j\|$ فاصله‌ی بین موقعیت‌های s_i و s_j است. پارامتر φ دامنه‌ی فضایی و پارامتر ν هموارساز میدان تصادفی است ([۷]). مدل سلسله‌مراتبی ذکرشده را مدل فرایند گاوسی (GP) مستقل می‌گویند.

فرض کنید $\theta = (\beta, \sigma_\varepsilon^2, \sigma_\eta^2, \varphi, \nu)$ بردار پارامترهای مدل، $\mathbf{z}_{\ell t}^{obs}$ بردار داده‌های مشاهده شده و $\mathbf{z}_{\ell t}^*$ بردار داده‌های گمشده باشد. در این مقاله برای برآورد پارامترها از رهیافت بیزی استفاده می‌شود. با فرض استقلال پیشین پارامترها، توزیع پیشین توأم $\pi(\theta)$ به صورت

$$\pi(\theta) = \pi(\beta) \pi(\sigma_\varepsilon^2) \pi(\sigma_\eta^2) \pi(\varphi) \pi(\nu)$$

به‌دست می‌آید و توزیع‌های پیشین متداول برای پارامترهای مدل را به صورت

$$\begin{aligned} \beta &\sim N_p(0, \sigma_\beta^2 I_p), & \sigma_\varepsilon^2 &\sim IG(a_\varepsilon, b_\varepsilon), & \sigma_\eta^2 &\sim IG(a_\eta, b_\eta), \\ \varphi &\sim \Gamma(a_\varphi, b_\varphi), & \nu &\sim \Gamma(a_\nu, b_\nu) \end{aligned}$$

در نظر گرفته‌ایم. که در آن σ_β^2, a, b ابرپارامترهای توزیع‌های پیشین هستند. بنا بر این لگاریتم توزیع پسین توأم پارامترها و داده‌های گمشده برای مدل GP به‌صورت زیر به‌دست می‌آید،

$$\begin{aligned} \log \pi(\theta, \mathbf{u}_{\ell t}, \mathbf{z}_{\ell t}^* | \mathbf{z}_{\ell t}^{obs}) &\propto -\frac{N}{\nu} \log \sigma_\varepsilon^2 \\ &- \frac{1}{\sigma_\varepsilon^2} \sum_{\ell=1}^r \sum_{t=1}^{T_\ell} (\mathbf{z}_{\ell t} - \mathbf{u}_{\ell t})' (\mathbf{z}_{\ell t} - \mathbf{u}_{\ell t}) - \frac{\sum_{\ell=1}^r T_\ell}{\nu} \log |\sigma_\eta^2 S_\eta| \\ &- \frac{1}{\sigma_\eta^2} \sum_{\ell=1}^r \sum_{t=1}^{T_\ell} (\mathbf{u}_{\ell t} - X_{\ell t} \beta)' S_\eta^{-1} (\mathbf{u}_{\ell t} - X_{\ell t} \beta) + \log \pi(\theta) \end{aligned}$$

با توجه به این که توزیع پسین شکل بسته‌ای ندارد، از روش‌های MCMC برای تولید نمونه استفاده می‌شود. بعد از همگرایی و کنار گذاشتن مقادیر سوخته‌ی نمونه‌های تولید شده از الگوریتم‌های MCMC، برآورد پارامترها با میانگین‌گیری از این نمونه‌ها حاصل می‌شوند.

۳-۲- مدل سلسله‌مراتبی اتورگرسیو

مدل سلسله‌مراتبی اتورگرسیو (AR) به صورت

$$(۱) \quad \mathbf{z}_{\ell t} = \mathbf{u}_{\ell t} + \varepsilon_{\ell t}, \mathbf{u}_{\ell t} = \rho \mathbf{u}_{\ell, t-1} + X_{\ell t} \boldsymbol{\beta} + \boldsymbol{\eta}_{\ell t}$$

معرفی شده است [۲۳]، که در آن $\varepsilon_{\ell t}$ و $\boldsymbol{\eta}_{\ell t}$ همانند مدل GP نرمال فرض می‌شوند و ρ پارامتر همبستگی زمانی با حوزه‌ی مقادیر (۱ و -۱) است. اگر $\rho = 0$ باشد، مدل GP حاصل می‌شود.

در این مدل مقادیر شروع $\mathbf{u}_{\ell}$ برای هر $\ell = 1, \dots, r$ از توزیع $N(\boldsymbol{\mu}_{\ell}, \sigma_{\ell}^2 S_{\ell})$ مشخص می‌شود، که در آن S_{ℓ} ماتریس همبستگی فضایی حاصل از تابع همبستگی ماترن با پارامترهای φ و ν همانند $\boldsymbol{\eta}_{\ell t}$ است. بنا بر این همه‌ی پارامترهای مدل به صورت

$$\boldsymbol{\theta} = (\boldsymbol{\beta}, \rho, \sigma_{\varepsilon}^2, \sigma_{\eta}^2, \varphi, \nu, \boldsymbol{\mu}_{\ell}, \sigma_{\ell}^2)$$

می‌باشند. با در نظر گرفتن پیشین توأم $\pi(\boldsymbol{\theta})$ مشابه مدل GP برای پارامترها با این تفاوت که

$$\rho \sim U(-1, 1), \quad \sigma_{\ell}^2 \sim IG(a_{\ell}, b_{\ell}), \quad \boldsymbol{\mu}_{\ell} \sim N(0, \sigma_{\mu}^2 I_p)$$

و با لگاریتم گرفتن، توزیع پسین توأم پارامترها و مقادیر گمشده به صورت

$$\log \pi(\boldsymbol{\theta}, \mathbf{u}_{\ell t}, \mathbf{z}_{\ell t}^* | \mathbf{z}_{\ell t}^{obs}) \propto -\frac{N}{2} \log \sigma_{\varepsilon}^2 - \frac{1}{2 \sigma_{\varepsilon}^2} \sum_{\ell=1}^r \sum_{t=1}^{T_{\ell}} (\mathbf{z}_{\ell t} - \mathbf{u}_{\ell t})' (\mathbf{z}_{\ell t} - \mathbf{u}_{\ell t}) - \frac{\sum_{\ell=1}^r T_{\ell}}{2} \log |\sigma_{\eta}^2 S_{\eta}|$$

$$-\frac{1}{2\sigma_\eta^2} \sum_{l=1}^r \sum_{t=1}^{T_\ell} (\mathbf{u}_{\ell t} - \rho \mathbf{u}_{\ell t-1} - X_{\ell t} \boldsymbol{\beta})' S_\eta^{-1} (\mathbf{u}_{\ell t} - \rho \mathbf{u}_{\ell t-1} - X_{\ell t} \boldsymbol{\beta})$$

$$-\frac{1}{2} \sum_{\ell=1}^r \log |\sigma_\ell^2 S_\circ| - \frac{1}{2\sigma_\ell^2} \sum_{\ell=1}^r \sum_{t=1}^{T_\ell} (\mathbf{u}_{\ell \circ} - \boldsymbol{\mu}_\ell)' S_\circ^{-1} (\mathbf{u}_{\ell \circ} - \boldsymbol{\mu}_\ell) + \log \pi(\boldsymbol{\theta}).$$

به دست می‌آید و این توزیع پسین شکل بسته‌ای ندارد، بنا بر این برای برآورد پارامترها از روش‌های MCMC استفاده می‌شود.

۳-۳- مدل سلسله‌مراتبی پویا

مدل پویا (DYN) برای داده‌های فضایی - زمانی به صورت

$$z(s, t) = \mathbf{x}'(s, t) \boldsymbol{\beta}_t + u_t(s, t) + \varepsilon(s, t), \quad \varepsilon(s, t) \sim N(\circ, \tau_t^2)$$

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim N(\circ, \Sigma_\eta)$$

$$u(s, t) = u(s, t-1) + w(s, t),$$

$$(2) \quad w(s, t) \sim GP\left(\circ, \sigma_t^2 \rho_t(\circ, \varphi_t)\right), \quad t = 1, \dots, N_t$$

ارایه شده است $([1^\circ])$ ، که در آن $GP\left(\circ, \sigma_t^2 \rho_t(\circ, \varphi_t)\right)$ یک فرایند گاوسی با تابع همبستگی فضایی $\rho_t(\circ, \varphi_t)$ است و σ_t^2 و φ_t پارامترهای تابع کوواریانس فضایی هستند. همچنین فرض می‌شود $\boldsymbol{\beta}_\circ \sim N(\mathbf{m}_\circ, \Sigma_\circ)$ و $u(s, \circ) = \circ$ است. با قرار دادن

$$\mathbf{z}_t = \left(z(s_1, t), \dots, z(s_n, t)\right)'$$

و X_t ماتریس $n \times p$ که i امین سطر آن $\mathbf{x}'(s_i, t)$ است. شکل ماتریسی مدل به صورت

$$\mathbf{z}_t = X_t \boldsymbol{\beta}_t + \mathbf{u}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N(\circ, \tau_t^2 I),$$

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \quad \mathbf{u}_t = \mathbf{u}_{t-1} + \mathbf{w}_t, \quad t = 1, \dots, N_t,$$

می‌شود که در آن $\mathbf{w}_t = (w(s_1, t), \dots, w(s_n, t))'$ و $\mathbf{u}_t = (u(s_1, t), \dots, u(s_n, t))'$ است. بنا بر این پارامترهای مدل به صورت

$$\theta'_t = (\beta'_\circ, \Sigma_\eta, \{\beta_t\}', \{\mathbf{w}_t\}', \{\tau_t^\gamma\}, \{\sigma_t^\gamma\}, \{\varphi(t)\})$$

می‌باشند. برای تحلیل بیزی این مدل با در نظر گرفتن پیشین‌های مستقل برای پارامترها، توزیع پسین توأم به شکل

$$\begin{aligned} \pi(\theta_t | \mathbf{z}_t) &\propto \prod_{t=1}^{N_t} N(\mathbf{z}_t | X_t \beta_t + \mathbf{u}_t, \tau_t^\gamma I_n) \prod_{t=1}^{N_t} \pi(\tau_t^\gamma) \\ &\prod_{t=1}^{N_t} N(\beta_t | \beta_{t-1}, \Sigma_\eta) N(\beta_\circ | \mathbf{m}_\circ, \Sigma_\circ) \pi(\Sigma_\eta) \\ (3) \quad &\prod_{t=1}^{N_t} \pi(\sigma_t^\gamma) \prod_{t=1}^{N_t} \pi(\mathbf{u}_t) \prod_{t=1}^{N_t} \pi(\mathbf{w}_t) \end{aligned}$$

به‌دست می‌آید که در آن توزیع پیشین پارامترها به صورت

$$\Sigma_\eta \sim W^{-1}(\sigma^\gamma I_p, \nu_\eta), \quad \mathbf{w}_t \sim N_n(\circ, \sigma_w^\gamma I_n), \quad \sigma_t^\gamma \sim IG(a_\sigma, b_\sigma)$$

تعیین شده‌اند و W^{-1} نماد توزیع ویشارت معکوس است. همان‌طور که ملاحظه می‌شود (۳) دارای شکل بسته‌ای نیست و همانند مدل‌های اتورگرسیو و فرایند تصادفی نیازمند استفاده از روش‌های MCMC است. با اجرای الگوریتم MCMC و تولید نمونه‌ی برآورد پارامترهای مدل محاسبه می‌شوند.

در ادامه این سه مدل روی داده‌های قیمت مسکن شهر تهران برازش و مورد مقایسه قرار می‌گیرند. همه‌ی محاسبات با به کار گرفتن نرم‌افزار R انجام شده است.

۴- تحلیل داده‌های قیمت مسکن

سه مدل سلسله‌مراتبی بیزی GP، AR و DYN با در نظر گرفتن پیشین‌های معمول معرفی شده برای پارامترهای مدل ([۱۱] و [۲۳])، روی داده‌های قیمت مسکن برازش شد. لازم به ذکر است که ابر پارامترهای توزیع پیشین طوری اختیار شده است که واریانس توزیع بزرگ باشد. با توجه به تحلیل اکتشافی صورت گرفته در بخش ۲، طول و عرض

جغرافیایی به‌عنوان متغیرهای کمکی و از لگاریتم قیمت مسکن به‌عنوان متغیر پاسخ استفاده شده است.

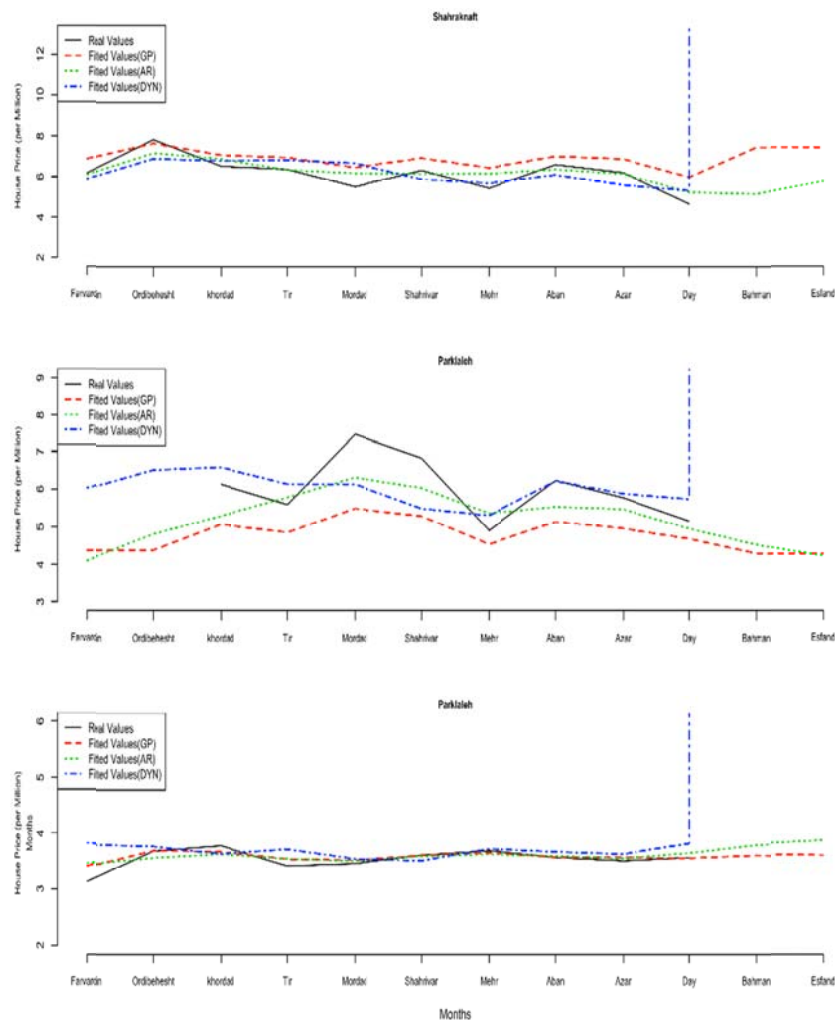
نتایج برازش مدل برای ۱۰۰۰۰ تکرار با مقادیر سوخته ۵۰۰۰ در جدول ۱ خلاصه شده‌اند. با توجه به نتایج جدول ۱ متغیر کمکی طول جغرافیایی تأثیر چندانی با احتمال قابل باور ۹۵ درصد روی قیمت مسکن ندارد، اما عرض جغرافیایی تأثیرگذار است. با توجه به ملاک مجذور میانگین توان دوم خطای پیش‌بینی و زمان اجرای مدل AR سلسله‌مراتبی برای داده‌های قیمت مسکن مناسب‌تر است. پارامترهای مدل DYN به زمان وابسته است و این باعث طولانی شدن زمان محاسبات شده است و به همین دلیل در جدول ۱ فقط نتایج ماه فروردین ارایه شده است.

جدول ۱- نتایج برازش بیزی سه مدل فضایی زمانی روی داده‌های قیمت مسکن

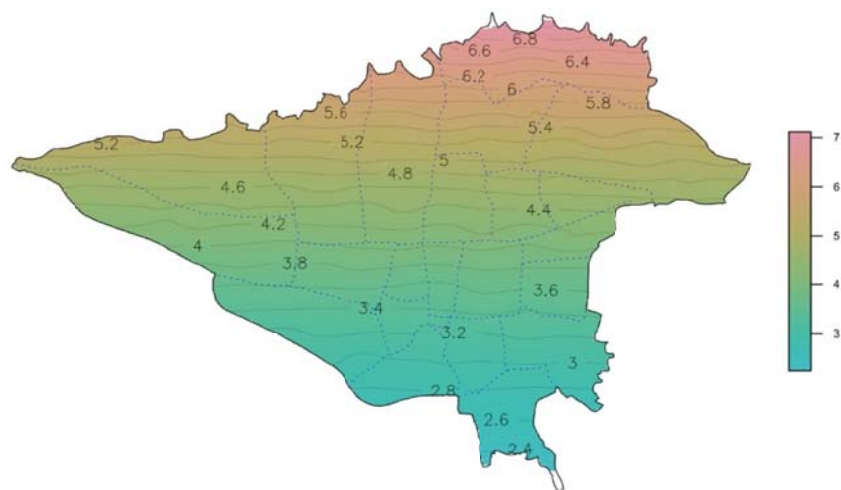
مدل DYN		مدل AR		مدل GP	
پارامتر	[فاصله اطمینان ۹۵٪] (انحراف معیار) برآورد	پارامتر	[فاصله اطمینان ۹۵٪] (انحراف معیار) برآورد	پارامتر	[فاصله اطمینان ۹۵٪] (انحراف معیار) برآورد
β_0	۱/۴۳۸۵، ۱/۴۶۱۳ [۱/۴۵۰۰ (۰/۰۵۸)]	$\beta_{t,t}$	۰/۸۲۳۳ (۰/۰۲۴۱) [۰/۷۹۵۵، ۰/۸۹۰۶]	β_0	۱/۴۵۰۰ (۰/۰۵۸)
β_1	۰/۲۶۱۱ (۰/۰۵۸) [۰/۲۴۹۵، ۰/۲۷۲۶]	$\beta_{t,t}$	۰/۱۵۱۱ (۰/۰۶۷) [۰/۱۳۸۳، ۰/۱۶۴۶]	β_1	۰/۲۶۱۱ (۰/۰۵۸)
β_2	۰/۱۵۱۱ (۰/۰۷۳) [۰/۰۷۳، ۰/۱۵۱]	$\beta_{t,t}$	۰/۰۲۱ (۰/۰۵۳) [۰/۰۰۰، ۰/۰۸۴]	β_2	۰/۰۳۹ (۰/۰۵۷)
σ_e^2	۰/۱۱۶۲ (۰/۰۶۷) [۰/۰۳۴، ۰/۱۳۰۱]	τ_e^2	۰/۸۹۲ (۰/۰۴۸) [۰/۰۸۰، ۰/۰۹۸۹]	σ_e^2	۰/۱۱۶۲ (۰/۰۶۷)
σ_η^2	۰/۱۰۷۴ (۰/۰۹۷۴) [۰/۰۹۷۴، ۰/۱۰۷۴]	σ_e^2	۰/۸۴۳ (۰/۰۲۱) [۰/۰۸۰، ۰/۰۸۸۶]	σ_η^2	۰/۱۰۷۴ (۰/۰۹۷۴)
ϕ	۱/۱۲۴۷ (۰/۰۸۳۷۸) [۰/۱۰۵، ۰/۳۶۳۶]	ϕ_t	۰/۴۲۰۲ (۰/۰۱۶۲) [۰/۳۸۸۶، ۰/۴۵۲۱]	ϕ	۱/۱۲۴۷ (۰/۰۸۳۷۸)
ρ	—	ρ	۰/۴۲۰۲ (۰/۰۱۶۲) [۰/۳۸۸۶، ۰/۴۵۲۱]	ρ	—
زمان اجرا	۴۰ دقیقه و ۳۸ ثانیه	زمان اجرا	۵۹ دقیقه و ۱۷ ثانیه	زمان اجرا	۴۰ دقیقه و ۳۸ ثانیه
RMSE	۱/۲۵۲۸	RMSE	۱/۱۰۱۷	RMSE	۱/۲۵۲۸

پیشگویی رفتار قیمت مسکن برای ۳ محله‌ی شهر تهران براساس سه مدل در شکل ۵ ارایه شده است که نشان می‌دهد مدل AR حتی با وجود داده‌ی گمشته نسبتاً زیاد بهتر از دو مدل دیگر برای این داده‌ها عمل می‌کند. یکی از مشکلات مدل DYN این است که برای پیش‌بینی در زمان‌های آینده که در هیچ موقعیتی مشاهده وجود ندارد عملکرد بدی را نشان می‌دهد که در شکل ۵ به‌وضوح دیده می‌شود. در انتها نقشه‌ی پیشگویی قیمت مسکن همه‌ی محله‌های شهر تهران بر اساس مدل AR برای ماه اسفند در شکل ۶ ارایه شده است که نحوه‌ی توزیع قیمت مسکن را بر اساس طیف‌های رنگی متفاوت روی نقشه تهران نمایش می‌دهد. در واقع نقاط هم‌رنگ تقریباً قیمت مسکن نوساز یکسانی دارند و از جنوب به شمال و شرق به غرب میزان تغییر قیمت را نشان می‌دهد. انحراف معیار پیشگویی در شکل ۷ رسم شده است که دقت پیشگویی را برحسب موقعیت

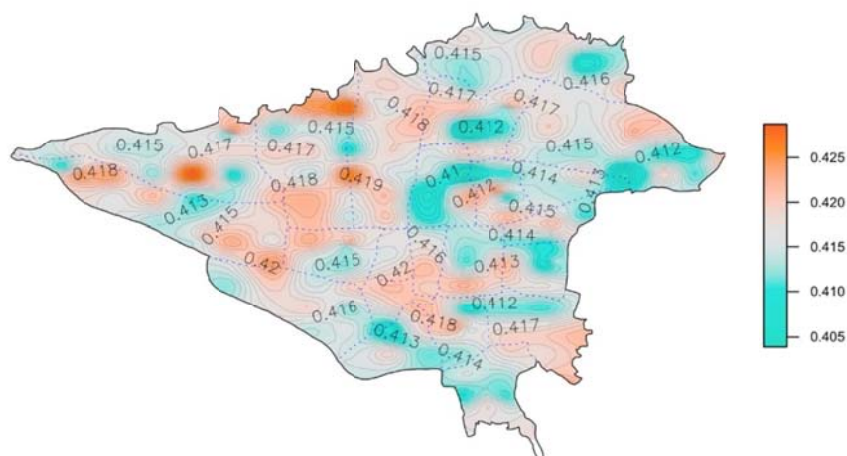
نشان می‌دهد. با توجه به این نقشه می‌توان نتیجه گرفت انحراف معیار نقاط پیشگویی در محله‌هایی که تفاوت قیمت مسکن در آن زیاد است، زیاد شده است.



شکل ۵- برازش سه مدل برای سه محله‌ی شهر تهران



شکل ۶- نقشه‌ی پیشگویی قیمت مسکن در شهر تهران برای ماه اسفند ۱۳۹۴ برحسب میلیون تومان به متر مربع



شکل ۷- انحراف معیار پیشگویی قیمت مسکن در شهر تهران در ماه اسفند ۱۳۹۴

۵- بحث و نتیجه‌گیری

در این مقاله سه مدل فضایی- زمانی برای داده‌های حجم قیمت مسکن در شهر تهران به‌کار گرفته شد. با استفاده از روش بیزی سلسله‌مراتبی برآورد پارامترهای مدل ارایه شد.

سپس این مدل‌ها را بر اساس ملاک RMSE و زمان اجرا مورد ارزیابی قرار دادیم. نتایج حاصل نشان دادند که مدل AR نسبت به مدل‌های دیگر برآزش نسبتاً مناسب‌تری به داده‌های قیمت مسکن دارند هر چند مدل DYN به نظر RMSE کم‌تری نسبت به این مدل داشت اما از نظر زمان اجرا و پیش‌بینی در زمان‌های آینده که در هیچ موقعیتی مشاهده‌ای وجود ندارد از عملکرد خوبی برخوردار نبود و پارامترهای این مدل با افزایش بعد زمان زیادتر می‌شوند. به خصوص برای داده‌های حجیم‌تر مثل داده‌های قیمت مسکن در سال ۱۳۹۵ که در حدود ۳۵۰ محله و به‌صورت روزانه ثبت شده است، پارامترهای مدل زیاد و مدل پویا در عمل ناکارآمد خواهد بود.

با استفاده از مدل AR به تحلیل داده‌های حجیم قیمت مسکن در شهر تهران پرداخته شد. تأثیر دو متغیر کمکی طول جغرافیایی و عرض جغرافیایی بر روی قیمت مسکن مورد بررسی قرار گرفت و با توجه به فاصله‌ی قابل‌باور بیزی ۹۵ درصد، طول جغرافیایی تأثیر چندانی روی قیمت مسکن نداشت، اما عرض جغرافیایی تأثیر زیادی دارد. در نهایت نقشه‌ی پیشگویی قیمت مسکن در ماه اسفند برای کل ناحیه‌ی شهر تهران رسم شد. به‌عنوان پیشنهاد می‌توان رهیافت بیزی تقریبی INLA که توسط [۲۱] معرفی شده است را برای این داده‌ها به کار گرفت و با مدل‌های دیگر مقایسه کرد. به علاوه داده‌های قیمت مسکن چوله به راست هستند می‌توان با توزیع‌های چوله مدل‌بندی کرد، البته با وجود زمان از پیچیدگی‌های خاصی برخوردار است.

توضیحات

1. Markov Chain Monte Carlo
2. Local Kriging
3. Sparse Matrix
4. Gaussian Prediction Process
5. Dynamic Model
6. Autoregressive
7. Gaussian Process

مرجع‌ها

- [1] Banerjee, S., Gelfand, A.E., Finley, A.O. and Sang, H. (2008). Gaussian Predictive Process Models for Large Spatial Data Sets. *Journal of the Royal Statistical Society Series*, **70**, 825-848.
- [2] Banerjee, S., Carlin, B.P. and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton.
- [3] Bivand, R.S., Rubio-Gomez, V. and Rue, H. (2015). Spatial Data Analysis with R-INLA with Some Extensions. *Journal of Statistical Software*, **63**, 1-31.
- [4] Cameletti, M., Ignaccolo, R. and Bande, S. (2011). Comparing Spatio-Temporal Models for Particulate Matter in Piemonte. *Environmetrics*, **22**, 985-996.
- [5] Cameletti, M., Lindgren, F., Simpson, D. and Rue, H. (2013). Spatio-Temporal Modeling of Particulate Matter Concentration through the SPDE Approach. *Advances in Statistical Analysis*, **97**, 109-131.
- [6] Carlin, B.P. and Louis, T.A. (2011). *Bayesian Methods for Data Analysis*. 3rd edition. Taylor & Francis.
- [7] Cressie, N.A.C. (1993). *Statistics for Spatial Data*. John Wiley & Sons, New York.
- [8] Cressie, N.A.C. and Wikle, C.K. (2011). *Statistics for Spatio-Temporal Data*. John Wiley & Sons, New York.
- [9] Finley, A.O., Sang, H., Banerjee, S. and Gelfand, A. (2009). Improving the Performance of Predictive Process Modeling for Large Datasets. *Computational Statistics and Data Analysis*, **53**, 2873-2884.
- [10] Finley, A., Banerjee, S. and Gelfand, A. (2012). Bayesian Dynamic Modeling for Large Space-Time Datasets Using Gaussian Predictive Processes. *Journal of Geographical Systems*, **14**, 29-47.

- [11] Gelfand, A., Banerjee, S. and Gamerman, D. (2005). Univariate and Multivariate Dynamic Spatial Modelling. *Environmetrics*, **16**, 465–479.
- [12] Gelfand, A.E. (2012). Hierarchical Modeling for Spatial Data Problems. *Spatial Statistics*, **1**, 30–39.
- [13] Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC.
- [14] Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall.
- [15] Hartman, L. and Hössjer, O. (2008). Fast Kriging of Large Data Sets with Gaussian Markov Random Fields. *Computational Statistics & Data Analysis*, **52**, 2331–2349.
- [16] Khana, D., Rossen, L.M., Hedegaard, H. and Warner, M. (2018). A Bayesian Spatial and Temporal Modeling Approach to Mapping Geographic Variation in Mortality Rates for Subnational Areas with R-inla. *Journal of Data Science*, **16**, 147–182.
- [17] Martins, T.G., Simpson, D. and Lindgren, F. (2013). HR Bayesian Computing with INLA: New Features. *Journal of Computational Statistics and Data Analysis*, **67**, 68–83.
- [18] Matérn, B. (1986). *Spatial Variation*. 2nd edition. Springer-Verlag, Berlin.
- [19] Robert, C. and Cassella, G. (2004). *Monte Carlo Statistical Methods*, Second edition, Springer, New York.
- [20] Rue, H. and Held, L. (2006). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC, Boca Raton.
- [21] Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian Inference for Latent Gaussian Models by using Integrated Laplace Approximation. *J. Roy. Statist. Soc. Ser.*, **71**, 319–392.

- [22] Sahu, S.K. and Bakar, K.S. (2011). A Comparison of Bayesian Models for Daily Ozone Concentration Levels. *Statistical Methodology*, **9**, 144-157.
- [23] Sahu, S.K. and Bakar, K.S. (2012). Hierarchical Bayesian Autoregressive Models for Large Space Time Data with Applications to Ozone Concentration Modelling. *Applied Stochastic Models in Business and Industry*, **28**, 395-415.
- [24] Sahu, S.K., Gelfand, A.E. and Holland, D.M. (2007). High-Resolution Space-Time Ozone Modeling for Assessing Trends. *Journal of the American Statistical Association*, **102**, 1221-1234.

امید کریمی

دکتری آمار

سمنان، دانشگاه سمنان، گروه آمار.

رایانشانی: omid.karimi@semnan.ac.ir

فاطمه حسینی

دکتری آمار

سمنان، دانشگاه سمنان، گروه آمار.

رایانشانی: fatemeh.hoseini@semnan.ac.ir