اولین کنفرانس مهندسی فناوری اطلاعات مکانی
دانشگاه صنعتی خواجه نصیرالدین طوسی
دانشکده مهندسی نقشه‌برداری
۲۹ و ۳۰ دی ماه ۱۳۹۴

The 1st National Conference on
Geospatial Information Technology
K.N.Toosi University of Technology
Faculty of Geomatics Engineering
19 - 20 January 2016

بررسی مدل‌سازی شبه‌پارامتری رگرسیون منطقی وزن‌دار جغرافیایی با شرایط متفاوت در وزن‌دهی مطالعه موردی بررسی تاثیر عوامل محیطی بر بیماری حساسیت

مرتضی دادگری[1*]،فرید کریمی‌پور [2] ،علی اسماعیلی [3]

۱- دانشجوی کارشناسی ارشد، سیستم اطلاعات جغرافیایی دانشکده عمران و نقشه برداری، دانشگاه تحصیلات تکمیلی صنعتی و فناوری پیشرفته کرمان
۲- استادیار، سنجش از دور، دانشکده عمران و نقشه برداری، دانشگاه تحصیلات تکمیلی صنعتی و فناوری پیشرفته کرمان

چکیده :

استفاده از خودهمبستگی مکانی و ناایستایی مکانی داده‌ها، در پژوهش‌های که با داده‌های مکانی انجام می‌شوند باعث شده که رگرسیون منطقی وزن‌دار جغرافیایی از جمله کاراترین روش‌های پردازش داده‌های مکانی در بین انواع مختلف رگرسیون‌های مکانی باشد. مدل‌سازی با این روش می‌تواند به دو صورت سراسری که شامل یک ضریب ثابت برای هر متغیر مستقل و شبه‌پارامتری انجام شود که مدل شبه-پارامتری خود شامل دو قسمت سراسری و محلی برای ضرایب هر متغیر، که در هر نقطه تغییر می‌کند، می‌باشد. از سوی دیگر، بیماری حساسیت از بیماری‌هایی است که به دلیل آلوده شدن محیط اطراف زندگی انسان به عوامل محرک حساسیت، گسترش فزاینده‌ای در سالیان اخیر داشته است. درمان یا تعدیل اثرات آلرژی با شناسایی عامل آلرژی‌زا و سپس، دوری جستن از آن عامل امکان‌پذیر است. در این میان، عوامل ناشناخته بسیاری هستند که می‌توانند موجب ایجاد آلرژی گردند. ایجاد نقشه مکانی خطرپذیری حساسیت می‌تواند برای بیماران در شناسایی مناطق پرخطر برای این بیماری مفید باشد. این مقاله، به بررسی وابستگی میان وقوع حساسیت با پارامترهای محیطی همچون فاصله از پارک و آلاینده‌های هوا مانند CO، SO_2، NO_2، PM_{10}، $PM_{2.5}$ و O_3 می‌پردازد. در انجام این پژوهش روش رگرسیون منطقی وزن‌دار جغرافیایی با شرایط متفاوت در هسته وزن‌دهی و نوع مدل‌سازی (سراسری یا محلی) استفاده شده است. نتایج نشان‌دهنده این است که پارامترهای فاصله از پارک و آلاینده‌های CO، NO_2، PM_{10} و $PM_{2.5}$ بر وقوع آلرژی تاثیرگذار هستند و آلاینده-های CO و O_3 تأثیری در وقوع و تشدید این بیماری ندارند. نتایج هنگامی که از روش مدل‌سازی شبه‌پارامتری با هسته وزندهی تطبیق پذیر بدست آیند نسبت به سایر روش‌ها قابل اعتمادتر هستند. در نهایت با استفاده از عوامل تاثیرگذار و روش رگرسیون منطقی وزن‌دار نقشه خطر شیوع حساسیت تولید شد.

واژه‌های کلیدی : خودهمبستگی مکانی، رگرسیون منطقی وزندار جغرافیایی،، بیماری حساسیت

نویسنده مکاتبه کننده : مرتضی دادگری
آدرس پستی : کرمان – شهرک مطهری- خیابان کرامت مجتمع راه و ترابری بلوک E – واحد ۳۲
تلفن : ۰۹۳۸۸۶۸۵۹۳۵
آدرس پست الکترونیک : morteza.dadgari@yahoo.com

۱- مقدمه

داده‌کاوی در دهه‌های اخیر پیشرفته چشم‌گیری داشته و فرآیند کشف دانش با طیف گسترده داده‌های امروزی از روش‌های آماری سنتی تقریبا غیر ممکن می‌نماید. با افزایش چشم‌گیر حجم این داده‌ها از روش‌های سابق بر این که بر گزاره‌های آماری معمولی متکی بودند پردازش‌ها امکان پذیرنمی‌باشد و به سمت روش‌های داده‌کاوی سوق داده شده‌اند[۱]. در این پژوهش تلاش شده یکی از روش‌های کاوش داده‌های مکانی، رگرسیون منطقی وزن‌دار جغرافیایی دقیق‌تر بررسی شود و نتایج حالات مختلف این مدل‌سازی با یکدیگر مقایسه گردد. در این راه از پژوهشی که با این روش بر روی داده‌های تعیین تاثیر گذاری عوامل محیطی مختلف بر بیماری حساسیت انجام گرفته، استفاده شده است.

امروز با رشد جوامع، بیماری‌ها نیز از حیث تنوع و تعداد افراد درگیر، در حال افزایش هستند. یکی از بیماری‌هایی که گسترش فزاینده‌ای داشته است، بیماری حساسیت یا حساسیت است [۱۲]. حساسیت واکنش افراطی سیستم ایمنی بدن نسبت به عوامل گوناگون است. کسانی که دچار حساسیت هستند، دارای دستگاه ایمنی فوق هوشیار هستند که نسبت به مواد ظاهراً بی ضرر موجود در محل زندگی‌شان، واکنشی بیش از حد معمول نشان می‌دهند. حساسیت یک مشکل بسیار شایع است و تقریباً از هر ده نفر، دو نفر به نوعی از آن مبتلا هستند. عوامل مختلفی در ایجاد و تشدید این بیماری دخیل هستند که بسته به نوع جغرافیا، شرایط محیطی و سبک زندگی افراد متغیر است. شناسایی عوامل محرک حساسیت و پیشگیری از مواجهه با مواد حساسیت‌زا بهترین راه برای جلوگیری از بروز حساسیت است[۱۲]. از آنجا که بخش مهمی از این عوامل، با محیط زندگی مرتبط است، کشف عوامل محیطی موثر بر شیوع حساسیت می‌تواند، نقش بسزایی در کاهش اثرات آن داشته باشد. از این رو می‌توان با جمع‌آوری اطلاعات مناسب از محیط زندگی افراد، نقش عوامل مختلف محیطی را در وقوع و تشدید این بیماری سنجید[۱۳]. هدف از مدل‌سازی و پردازش داده‌ها تشخیص عوامل موثر بر بیماری حساسیت در شهر تهران از میان عوامل انتخاب شده می‌باشد. این عوامل شامل آلاینده‌های هوا CO، SO_2، NO_2، PM_{10}، $PM_{2.5}$ و O_3 و فاصله از فضای سبز در شهر تهران می‌باشند که از تحلیل رگرسیون منطقی وزندار جغرافیایی در شرایط مدل‌سازی متفاوت ارزیابی می‌شوند.

از جمله پژوهش‌های که پیش از این حول این موضوع صورت گرفته شده می‌توان به پژوهشی که در سال ۲۰۱۲، Youssef Agha و همکاران [۷] از قابلیت‌های داده‌کاوی برای پیش‌بینی بروز حساسیت در بین دانش‌آموزان مدرسه ابتدایی بهره بردند، اشاره کرد. هدف از این تحقیق، پی بردن به رابطه میان وقوع حساسیت در بچه‌های مدرسه ابتدایی و مشاهدات روزانه هواشناسی بالای سطح زمین (دما، رطوبت نسبی، نقطه شبنم) و آلودگی هوای روزانه (ریزگردها، سولفور دی‌اکسید، نیتروژن دی‌اکسید، کربن مونواکسید و اوزون) با استفاده از رگرسیون منطقی بود. نتایج این تحقیق بیانگر آن است که افزایش دمای هوا و هم‌چنین غلظت بالای آلاینده‌های SO_2، CO و PM_{10} وابستگی زیادی با وقوع و تشدید بیماری حساسیت دارند. و همچنین از دیگر پژوهش‌ها می‌توان به تجزیه تحلیل رگرسیون جغرافیایی وزندار برای تشخیص سرطان پرستات در مراحل آخر در فلوریدا که توسط P Goovaerts و همکارانش در سال ۲۰۱۵ صورت گرفت اشاره نمود[۸]. آنها در طول دوره ۲۰۰۱ تا ۲۰۰۷ بیماران سرطان پرستات را مورد ارزیابی و با استفاده از رگرسیون منطقی مکانی وزن‌دار در محوطه‌ای به شعاع ۱۲۵ کیلومتر از فلوریدا را مورد مطالعه قرار دادند. از جمله عواملی که داده‌های آن از سرشماری جمع آوری شد، شامل خصوصیات افراد مبتلا به بیماری سرطان پرستات از قبیل سفید یا سیاه بودن رنگ پوست، روستایی یا شهری بودن، مصرف سیگار، وضعیت تاهل و سطح درآمد خانواده می‌توان نام برد. در نهایت رگرسیون وزن‌دار جغرافیایی به آنها اجازه داد که بتوانند تقسیم‌بندی از این ایالت ارائه دهند که نشان دهنده مکان‌های با بیشترین احتمال مبتلا شدن به این بیماری تا مکان‌های کم خطرتر باشد.

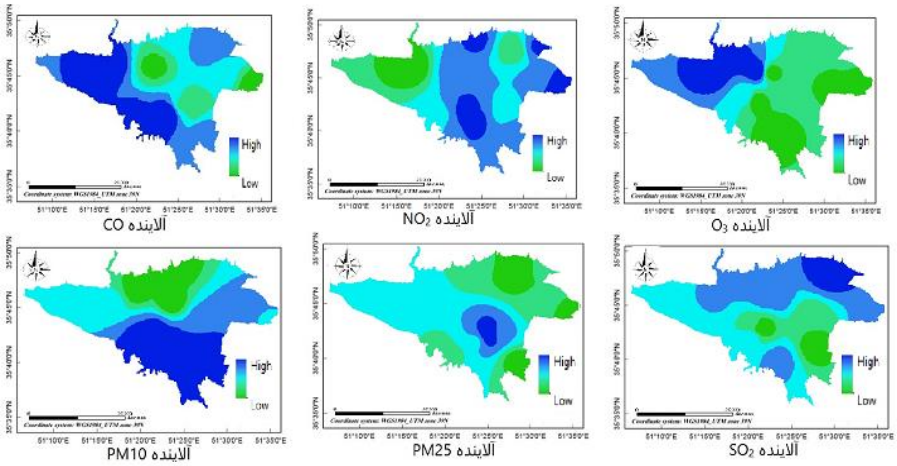
ساختار ادامه مقاله بدین شرح است: بخش دوم، به معرفی ویژگی‌های منحصر بفرد داده‌های این پژوهش را بیان می‌کند. در بخش سوم، رگرسیون وزن‌دار جغرافیایی معرفی می‌شود. در بخش چهارم، رگرسیون وزن‌دار جغرافیایی

شبه‌پارامتری ارزیابی شده و در بخش پنجم مراحل پیاده‌سازی انجام شده بر روی داده‌ها را تشریح نموده و نتایج حاصل از آن را ارائه و ارزیابی خواهیم نمود. در نهایت، در بخش ششم، نتیجه‌گیری حاصل از تحقیق ارائه خواهد شد.

۲- ویژگی‌های منحصر بفرد داده‌های این پژوهش

در تحلیل داده‌های مربوط به این پژوهش، با ویژگی‌های منحصر به فردی رو به رو هستیم که از آن جمله می‌توان به خودهمبستگی مکانی[1] و ناایستایی مکانی[2] اشاره کرد. به دلیل چنین خصوصیاتی لازم است در تحلیل داده‌های مکانی و کاوش در پایگاه‌هایی که همه یا بخشی از داده‌های آن را داده‌های مکان مرجع تشکیل می‌دهد، دقت بیشتری داشته باشیم [۱ و ۲]. خودهمبستگی مکانی و ناایستایی مکانی ویژگی‌هایی هستند که به ترتیب به داده‌های مکانی و پارامترهای مدل مکانی مربوط می‌شوند. در صورتی که چنین ویژگی‌هایی وجود داشته باشند ولی در برآورد پارامترهای مدل به حساب نیاید، ممکن است به مدلی ضعیف و بنابراین پیش‌بینی و تفسیر مکانی نامناسبی از داده‌ها، منجر شوند [۳]. وجود خودهمبستگی مکانی که ویژگی منحصر به فرد داده‌های مکان مرجع می‌باشد، استقلال داده‌ها از یکدیگر را که پیش‌فرض بسیاری از روش‌های آماری مرسوم است، برنمی‌تابد و با آن مغایرت دارد [۵]. از دیدگاه مدل‌سازی، ایستایی مکانی فرآیندی را توصیف می‌نماید که پارامترهای مدل بازگوکننده آن فرآیند، در تمام ناحیه مورد مطالعه سازگار باشند و در صورتی که پارامترهای مدل دارای ناسازگاری مکانی هستند، یعنی به طور محلی تغییر می‌کنند، انتخاب یک مدل سراسری منجر به استنباط مکانی نادرستی از رفتار پدیده مورد مطالعه می‌شود. در چنین شرایطی نمی‌توان نتایج به دست آمده از زیرمجموعه‌ای از کل داده‌ها را با نتایج به دست آمده از کل داده‌ها یکسان دانست [۴].

از سوی دیگر آلودگی هوا در شهر تهران با ایستگاه‌های کنترل آلودگی توسط شهرداری تهران و سازمان محیط زیست اندازه‌گیری می‌شود. پارامترهای آلودگی هوا مورد مطالعه در پژوهش، غلظت آلاینده‌های CO، SO_2، NO_2، PM_{10}، $PM_{2.5}$ و O_3 هستند. این داده‌ها که به صورت ساعتی در هر ایستگاه، اندازه‌گیری شده‌اند و برای بازه‌های زمانی ماه‌های خرداد و آذر سال ۹۲ از شهرداری تهران و سازمان محیط زیست اخذ گردیدند و به منظور رفع داده‌های افزونه و کاهش حجم زیاد داده‌ها به پارامتر آلودگی ماهانه، برای هر آلاینده، بیشینه غلظت آلودگی برای یک روز به عنوان آلودگی روزانه در نظر گرفته شد و با میانگین‌گیری از غلظت‌های روزانه برای هر ماه، آلودگی متوسط ماهانه برای هر ایستگاه محاسبه گردید.این مقادیر که میزان غلظت هر آلاینده را در ایستگاه‌های سنجش آلودگی نشان می‌دهند، برای تولید نقشه آلودگی ماهانه هر آلاینده از روش درون‌یابی مکانی کریجینگ استفاده شده است. برای هر آلاینده نقشه آلودگی متوسط ماه‌های خرداد و آذر به صورت پیوسته در کل شهر تولید شد (شکل (۱)).



شکل ۱: نقشه آلاینده‌های هوا برای شهر تهران

[1] Spatial Autocorrelation
[2] Spatial Non-Stationary

۳- مدل‌سازی رگرسیون وزن‌دار جغرافیایی

این مدل کارایی بالاتری نسبت به مدل رگرسیون سنتی دارد و با ضریبی که وارد می‌کنیم می‌توانیم قرائت‌های جدیدی از تغییر پدیده‌ها داشته باشیم. برخی از متغیرها را می‌توان برای تمام مکان‌ها با یک وزن وارد کرد و برخی دیگر را باید با تغییر مکان ضریب را نیز تغییر داد که با این گفته معادله سنتی رگرسیون به صورت کامل‌تر به شکل زیر در می‌آید که به آن صورت شبه‌پارامتری می‌گوییم.

$$y_i = \sum_k \beta_k(u_i, v_i)x_{k,i} + \sum_l \gamma_l z_{l,i} + \varepsilon_i \qquad (۱)$$

در معادله بالا $Z_{l,i}$ مقدار lامین متغیر با ضریب ثابت γ_l است. β_k در این معادله ضریبی است که به صورت محلی تغییر می‌کند و قسم محلی را می‌سازد و γ_l ضریبی است که برای هر متغیر مقدار ثابتی را دارد و قسم سراسری را شامل می‌شود. مدل رگرسیون منطقی وزن‌دار جغرافیایی است که از معادلات (۲ و ۳) برای مدل‌سازی استفاده می‌کند و نتایج را از این معدلات برای هر نقطه به دست می‌آورد. [۱۰].

$$y_i \sim Bernoulli\,[P_i]$$

$$logit(P_i) = \sum_k \beta_k(u_i, v_i)x_{k,i} \qquad (۲)$$

و مدل شبه‌پارامتری به صورت زیر است.

$$y_i \sim Bernoulli\,[P_i]$$

$$logit(P_i) = \sum_k \beta_k(u_i, v_i)x_{k,i} + \sum_l \gamma_l z_{l,i} \qquad (۳)$$

نکته مهم در (۳)، انتخاب مناسب پهنای باند است. به گونه‌ای که رشد پهنای باند به سمت بینهایت، راه‌حل مساله را به جواب حاصل از رگرسیون کمترین مربعات معمولی نزدیک می‌کند و با کوچک شدن پهنای باند؛ برآورد پارامترها به موقعیت‌های نزدیک به i وابسته شده و در نتیجه واریانس افزایش می‌یابد. به طور کلی دو گزینه در انتخاب پهنای باند هسته تابع وزن وجود دارد. انتخاب یک هسته مکانی با پهنای باند ثابت[۳] و انتخاب یک هسته مکانی با پهنای باند تطبیق‌پذیر[۴] که هسته‌های مکانی با پهنای باند تطبیق‌پذیر در رگرسیون وزندار جغرافیایی این قابلیت را دارند که بر اساس چگالی موقعیت‌های مشاهداتی و اندازه تغییرات، خود را تطبیق دهند. به این ترتیب برای موقعیت‌هایی که تراکم مشاهدات در همسایگی‌های آنها بالا است، عرض باند کاهش یافته و برای موقعیت‌هایی که مشاهدات به صورت پراکنده از آنها قرار دارند، عرض باند افزایش می‌یابد. در نتیجه با کاهش پهنای باند هسته مکانی در نواحی‌ای که موقعیت‌های مشاهدات تراکم بالایی دارند، اثر واریانس کاهش می‌یابد و با افزایش پهنای باند هسته مکانی در نواحی‌ای که موقعیت مشاهدات پراکنده است، امکان برآورد نشدن احتمالی بعضی از ضرایب از بین می‌رود.

۴- پیاده‌سازی و نتایج

بخش پیاده‌سازی در این پژوهش در دو مرحله انجام پذیرفته است. اول مقایسه روش‌های مختلف ارزیابی رگرسیون منطقی وزن‌دار جغرافیایی و انتخاب مدل برای پردازش داده‌ها و پس از آن پردازش داده‌ها از روش پیشنهادی و تعیین نوع وابستگی عوامل محیطی به بیماری حساسیت.

[3] Spatial Fixed Kernel Bandwidth
[4] Spatial adaptive Kernel Bandwidth

۴-۱- بررسی نتایج مدل‌های مختلف و انتخاب روش قابل اعتمادتر

به طوری که پیش از این مشخص شده در این پژوهش متغیر مستقل بیماری حساسیت است و هفت متغیر مستقل اثر پارک و غلظت آلاینده‌های O_3، CO، $PM_{2.5}$، PM_{10}، NO_2 و SO_2 مورد مطالعه قرار می‌گیرند. مقادیر محاسبه شده ضرایب این متغیرها در جداول زیر آمده است.

جدول ۱: مقادیر متوسط محاسبه شده پارامترها برای مدل‌سازی های مختلف

متغیرها	سراسری با هسته وزن‌دهی ثابت و تطبیق‌پذیر	شبه‌پارامتری با هسته وزن-دهی ثابت	شبه‌پارامتری با هسته وزن‌دهی تطبیق‌پذیر
park_dist	۰.۱۱۹۶	۰.۱۴۹	۰.۱۹۵
O3	۰.۵۷۰۷	۰.۵۵۶	۰.۲۷۸
PM2.5	۰.۷۳۱۶	۰.۷۰۴	۰.۶۵۷
PM10	۰.۵۹۰۳	۰.۵۶۲	۰.۴۸۱
NO2	۰.۹۸۸۴	۰.۹۸۵	۰.۹۹۸
SO2	۰.۲۲۷۸	۰.۲۳۵	۰.۱۶۸
CO	۰.۵۱۵۲	۰.۵۱	۰.۳۷

با نگاه به جدول (۱) مشاهده می‌شود که مقادیر محاسبه شده برای هر ضریب با روش‌های متفاوت، یکسان نیست. تحلیل آن که کدام دسته از داده‌ها قابل اعتمادتر هستند را با نگاه به کل مدل‌سازی در جداول (۲) می‌توان ارزیابی شود. پارامترهای آماری پس از پردازش و بدست آوردن ضرایب هر متغیر به صورت سراسری و ثابت و یا شبه‌پارامتری، با محاسبه وزن ضرایب با پهنای باند ثابت و تطبیق‌پذیر در جدول‌های زیر آمده است.

جدول ۲: پارامترهای آماری مدل‌های استفاده شده

هسته وزن‌دهی	مدل	درجه آزادی/ناریبی	درجه آزادی	ناریبی
هسته وزن‌دهی ثابت	**مدل سراسری**	۰.۷۶۵	۳۹۸.۰۰	۳۰۴.۴۸۷
	مدل شبه‌پارامتری	۰.۷۵۵	۳۹۳.۴۶۷	۲۹۷.۱۸۴
	تفاوت	۱.۶۱۱	۴.۵۳۳	۷.۳۰۴
هسته وزن‌دهی تطبیق‌پذیر	**مدل سراسری**	۰.۷۶۵	۳۹۸.۰۰	۳۰۴.۴۸۷
	مدل شبه‌پارامتری	۰.۷۴۱	۳۹۰.۵۲۰	۲۸۹.۴۵
	تفاوت	۲.۰۱	۷.۴۸	۱۵.۰۳۸

آن گونه از جدول (۲) استنباط می‌شود مدل سراسری همان گونه که انتظار می‌رفت، تفاوتی را نداشته و تنها در مدل شبه‌پارامتری نهایی تفاوت‌های ایجاد شده، بنابراین فقط از بخش محلی متغیرها این تفاوت‌ها حاصل شده است. همچنین در روش تطبیق‌پذیر میزان بهبودی در مدل شبه‌پارامتری بیش از روش قبل است. مقایسه سطرهای تفاوت در جدول (۲) نمایانگر این موضوع است که ناریبی در حالتی که هسته وزن‌دهی ثابت بوده با استفاده از مدل شبه-پارامتری فقط به میزان ۷.۳۰۴ بهبود داشته که در مقایسه با هسته وزن‌دهی تطبیق‌پذیر که بیش از دو برابر این بهبودی اتفاق افتاده است برابر ۱۵.۰۳۸ می‌باشد و در کل می‌توان نتیجه گرفت که مدل‌سازی با متغیرهای شبه-پارامتری نسبت به حالتی که ضرایب آنها در فضای معادلات ثابت و سراسری هستند بهبود پیدا می‌کند که این بهبودی در نتایج با استفاده از روش وزن‌دهی تطبیق‌پذیر نسبت به روش ثابت بیشتر است.

۴-۲- پردازش داده‌های حساسیت با روش انتخاب شده

مشخص شد که برای بهبود نتایج بهتر است از مدل‌سازی شبه‌پارامتری رگرسیون منطقی وزن‌دار جغرافیایی با هسته وزن‌دهی تطبیق پذیر استفاده شود. با این روش یک بار برای ارزیابی مدل‌ها متغیرها محاسبه شدند که مقادیر متوسط آنها در جدول (۱) آورده شده است. اکنون باید با استفاده از گزاره‌های آماری تعیین کنیم که کدام عوامل محیطی در نظر گرفته شده بر بیماری حساسیت تاثیر گذار است که در این راه از پارامتر P-Value و همچنین مقایسه آماره T با و توزیع t-Student استفاده شده است.

جدول ۳: مقادیر متوسط متغیرها و پارامترهای مقایسه

متغیر	مقدار متوسط β	P-Value	میانگین آزمون T	T80%	T95%	بازه آزمون T
park_dist	۰.۱۵۲	۰.۰۰۰	۰.۵۷۵	۰.۸۴۲	۱.۶۵	[−0.6,0.6]
O3	۰.۵۰۷	۰.۳۷۹	۱.۷۲۱	۰.۸۴۲	۱.۶۵	[−1.7,1.7]
PM2.5	۰.۶۴۵	۰.۰۰۰	۰.۳۴۲	۰.۸۴۲	۱.۶۵	[−0.3,0.3]
PM10	۰.۴۷۲	۰.۰۴۲	۰.۹۱۲	۰.۸۴۲	۱.۶۵	[−0.9,09]
NO2	۰.۹۷	۰.۰۰۰	۰.۶۰۴	۰.۸۴۲	۱.۶۵	[−0.6,0.6]
SO2	۰.۱۲۷	۰.۳۴۸	۱.۵۱۹	۰.۸۴۲	۱.۶۵	[−1.5,1.5]
CO	۰.۵۳۱	۰.۰۰۶	۰.۱۶۲	۰.۸۴۲	۱.۶۵	[−0.1,0.1]

اگر مقدار P-Value نزدیک به صفر (زیر ۰.۰۵) باشد با درست انگاشتن فرض صفر (ضریب رگرسیون برابر با صفر) داده-ها دارای ناسازگاری هستند. به عبارت دیگر با به صفر گراییدن این مقدار این موضوع که ضریب رگرسیون غیرصفر است تأیید می‌شود و غیرصفر بودن این ضریب یعنی نقش‌پذیری متغیر مربوط به آن ضریب، در ساخت معادله رگرسیون و در واقع تأیید تأثیر این متغیر است[۴]. از جدول (۳) مشخص می‌شود که به ترتیب O_3 و SO_2 با مقادیر ۰.۳۷۹ و ۰.۳۴۸ فاصله قابل توجه‌ای با صفر دارند و این به معنی این است که با فرض صفر بودن مقدار تخمین آنها داده‌ها دارای ناسازگاری نمی‌شوند یا به دیگر سخن این دو عامل می‌توانند نامزد انتخاب حذف از عوامل تأثیرگذار باشند که برای تصمیم‌گیری نهایی نیاز به بررسی آزمون T داریم. از مقایسه آماره T و توزیع t-Student با درصد اطمینان مشخص می‌توانیم به معنی‌دار بودن ارتباط بین متغیر مستقل انتخاب شده و متغیر وابسته پی ببریم. قرار گرفتن آماره $t_{(k)}$ در بازه مشخص شده ($|T_{(k)}| \geq t_{(n-2,\frac{\alpha}{2})}$) به معنی پذیرش فرض صفر آزمون مبنی بر عدم وجود ارتباط معنی‌دار می‌باشد. در مقابل اگر این آماره در خارج از بازه قرار گیرد به این معنی است که یک ارتباط معنی‌دار بین متغیر پیش-بین و پاسخ وجود دارد. حال اگر این آماره در طرف مثبت بازه باشد این رابطه معنی‌دار مثبت است و اگر منفی که دارای یک ارتباط منفی هستیم.

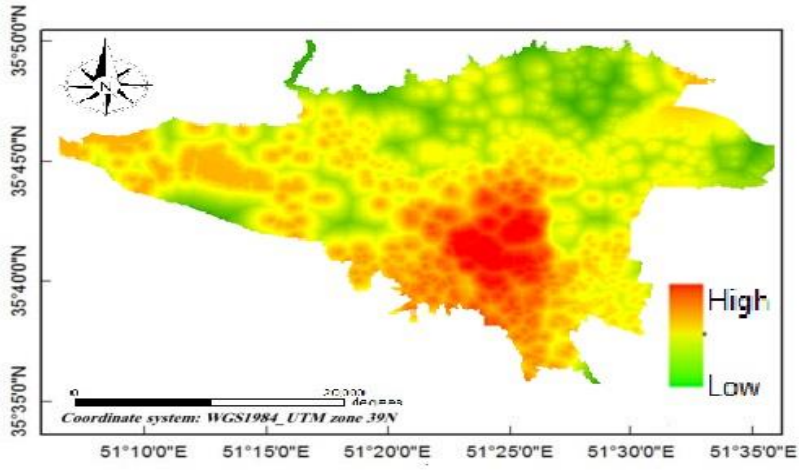
در این پژوهش دو سطح اطمینان ۸۰٪ و ۹۵٪ ($\alpha = 20\%$ و $\alpha = 5\%$) در نظر گرفته شده است که از جدول توزیع t-Student مقادیر با درجه آزادی ۴۱۰ (در واقع همان بی‌نهایت) مقادیر آن استخراج شد. حال اگر بازه محدوده آماره T را در نظر بگیریم (ستون آخر جدول (۳)) می‌توانیم ارزیابی نهایی را انجام دهیم در مورد متغیر O_3 هیچ یک از مقادیر با سطح اطمینان ۸۰٪ و ۹۵٪ در بازه قرار نگرفته‌اند و این نشان دهنده بی معنی بودن ارتباط O_3 با مدل است. از سوی دیگر در مورد SO_2 نیز مشاهده می‌شود که اگرچه در سطح اطمینان ۸۰٪ این رابطه معنی‌دار است اما در سطح ۹۵٪ این ارتباط نقض می‌شود و PM_{10} نیز حالت مشابه‌ای با SO_2 دارد. در نهایت دو نامزد حذف شدن نهایی

SO_2 و O_3 هستند و چون P-value در مورد PM_{10} ناسازگاری در ساختار مدل پیدا نکرده هم‌چنان در جمع عوامل تأثیرگذار (حتی با وجود آن‌که در سطح اطمینان ۹۵٪ قرار نمی‌گیرد) می‌ماند.

پس از شناسایی عوامل تاثیر گذار باید مدل‌سازی را با این عوامل انجام داده و نتیجه نهایی حاصل شود که بعد از انجام این پردازش معادله (۵) به دست آمد که در این معادله ضرایب متغیرها مقادیر متوسطی هستند که از روش رگرسیون منطقی وزن‌دار جغرافیای با هسته وزن‌دهی تطبیق پذیر حاصل شده‌اند.

$$Riskmap = Allergy = y = 0.285 * Parkdist + 0.046 * PM25 + 0.011 * PM10 + 0.035 * NO_2 + 0.032 * CO \quad (۵)$$

در نهایت با توجه به این مدل سازی می‌توان نقشه پهنه‌بندی خطر ابتلا یا تشدید بیماری حساسیت را تولید کرد که می‌توان از آن برای استفاده بیماران مبتلا در جهت پیش‌گیری از تشدید بیماری استفاده کرد.



شکل ۲: نقشه پهنه‌بندی خطر ابتلا یا تشدید بیماری حساسیت شهر تهران

۵- نتیجه‌گیری

در این مقاله به بررسی وابستگی میان وقوع حساسیت با عوامل محیط زندگی پرداخته شد. برای استخراج وابستگی‌های ناشناخته بین وقوع بیماری حساسیت با پارامترهای محیطی همچون فاصله از پارک و آلاینده‌های هوا (CO، SO_2، NO_2، PM_{10}، $PM_{2.5}$ و O_3)، از رگرسیون منطقی وزن‌دار جغرافیایی استفاده گردید که مشخص شد مدل‌سازی شبه-پارامتری با هسته وزن‌دهی تطبیق‌پذیر روشی کارا تر است. نتایج به دست آمده بیانگر این واقعیت هستند که پارامترهای میزان آلاینده‌های SO_2 و O_3 تاثیر معناداری در وقوع و تشدید حساسیت ندارد، در حالیکه سایر عوامل بررسی شده، باعث تشدید این بیماری می‌شوند. با تحلیل پارامتر P-Value و مقایسه آماره T و توزیع t-Student به دست آمده مشاهده می‌شود که جاهایی که به پارک نزدیک هستند و همچنین آلودگی خیلی زیادی دارند، مکان‌های مستعدتری برای شیوع بیماری حساسیت هستند. بنابراین برای درمان حساسیت و عدم تشدید این بیماری، بیمار باید از قرارگیری در این دست مکان‌ها دوری کند. همچنین با استفاده از عوامل موثر در تشدید حساسیت در این تحقیق مدلسازی مکانی خطر شیوع بیماری حساسیت صورت گرفت، که این مدل، بیمار را از مناطق پرخطر آگاه می‌سازد و بیمار می‌تواند با پرهیز از قرارگیری در این مکان‌ها به بهبود بیماری حساسیت خود کمک کند.

۵- منابع

[1] Cliff, A. and K. Ord, Testing for Spatial Autocorrelation Among Regression Residuals. Geographical Analysis, 1972. 4(3): p. 267-284.

[2] Fotheringham, A.S., C. Brunsdon, and M. Charlton, Geographically weighted regression: the analysis of spatially varying relationships. Vol: 2002 .13 : Wiley Chichester.

[3] Legendre, P., Spatial autocorrelation: trouble or new paradigm? Ecology, 1993. 74(6): p. 16591673.

[4] Brunsdon, C., A.S. Fotheringham, and M.E. Charlton, Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. Geographical Analysis, 1996. 28(4): p. 281-298.

[5] Fotheringham, A.S., C. Brunsdon, and M. Charlton, Geographically weighted regression: the analysis of spatially varying relationships2003: Wiley.

[6] Chatterjee, S. and A.S. Hadi, Regression analysis by example2013: John Wiley & Sons

[7] Searle, S.R., Linear models. 2012: John Wiley & Sons.

[8] Agresti, A., Categorical data analysis. 2014: John Wiley & Sons.

[9] Hosmer Jr, D.W., S. Lemeshow, and R.X. Sturdivant, Applied logistic regression. 2013: John Wiley & Sons.

[10] Wu, W. and L. Zhang, Comparison of spatial and non-spatial logistic regression models for modeling the occurrence of cloud cover in north-eastern Puerto Rico. Applied Geography, 2013. 37: p. 52-62.

[11] Harrell, F.E., Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. 2001: Springer.

[12] UK, A.; Available from: https://www.allergyuk.org/allergy-statistics/allergy-statistics.

[13]Asher, M., et al., International Study of Asthma and Allergies in Childhood (ISAAC): rationale and methods. European respiratory journal, 1995. 8(3): p. 483-491.

Investigating geographical weighted logistic regression Semi-parametric modeling with different weighting condition; environmental factors effects on allergic diseases case study

Dadgari, M. *[1], Esmaili, A. [2]

1-Ms.c student of Geographic information system in Department of Civil engineering, Graduate University of Advanced Technology Kerman

2- Assisstant professor in Department of remote sensing, Department of Civil engineering, Graduate University of Advanced Technology Kerman

Abstract

Using spatial autocorrelation and non-stationarity of spatial data in researchs that deal with spatial data, have made geographically weighted logistic regression as an efficient method for geo-processing between varieties of spatial regressions. Using these methods, modeling can be implemented in two ways, global model that consists of a fixed coefficient for each independent variable and semi-parametric model that includes two parts which are global and local coefficients for each variable that change at any point of model. On the other hand Allergy have been spreaded extensively in recent years, because of the increase of air pollution and contamination of the surrounding environment of human life. Treating or alleviating the effects of allergy is possible by identifying the allergens and avoiding them. However, many unknown factors may cause allergic reactions. Producing a risk-allergy location map could be useful for patients enabling them to identify high risk areas. The present study examines the association between allergy and environmental parameters such as the distance to parks and air pollutants including CO, SO2, NO2, PM10, PM2.5 and O3. **U**sed in this study is geographical weighted logistic regression with different conditions, weighting kernel and modeling (global or local). The results indicate that the distance to parks and pollutants CO, NO2, PM10 and PM2.5 parameters effect on the occurrence of allergy and pollutants such as CO and O3 does not. Besides, When we use the semi-parametric modeling method by adaptive weighting kernel, obtained results are more trustworthy than other methods. Finally, using weighted logistic regression method and effective factors, allergy risk map is produced.

Keywords: autocorrelation, geographically weighted logistic regression, Allergy.