بسم تعالی

دانشگاه صنعتی اصفهان

پروژه کارشناسی

پریناز احمدی تشنیزی

شماره داشجویی : 9917893

استاد راهنما:

دکتر مریم کلکین نما

موضوع :

درباره تحلیل داده های مکانی

مقدمه

با گسترش فناوری‌های دیجیتال و افزایش حجم داده‌های تولید شده در حوزه‌های مختلف، تحلیل داده‌های مکانی به یکی از حوزه‌های مهم و پرکاربرد در علوم جغرافیایی و سیستم‌های اطلاعات جغرافیایی تبدیل شده است. داده‌های مکانی نه تنها شامل اطلاعات جغرافیایی هستند، بلکه می‌توانند شامل داده‌های محیطی، اجتماعی، اقتصادی و حتی سیاسی باشند. این داده‌ها به دلیل ماهیت چندبعدی و پیچیده‌شان، نیازمند روش‌های تحلیلی پیشرفته‌ای هستند که بتوانند ارتباطات و الگوهای پنهان در آن‌ها را کشف کنند.

در سال‌های اخیر، استفاده از داده‌های بزرگ در تحلیل‌های مکانی به شدت افزایش یافته است. این داده‌ها می‌توانند از منابع مختلفی مانند ماهواره‌ها، سنسورهای محیطی، شبکه‌های اجتماعی و حتی دستگاه‌های هوشمند جمع‌آوری شوند. تحلیل این داده‌ها نه تنها به درک بهتر از محیط زیست و تغییرات آن کمک می‌کند، بلکه می‌تواند در مدیریت منابع طبیعی، برنامه‌ریزی شهری و حتی پیش‌بینی بلایای طبیعی نیز نقش کلیدی ایفا کند.

هدف این تحقیق، بررسی روش‌های تحلیل داده‌های مکانی در سیستم‌های اطلاعات جغرافیایی و ارائه یک رویکرد کارآمد برای شناسایی الگوهای مکانی با استفاده از تکنیک‌های آماری و یادگیری ماشین است.

در این پژوهش، از ترکیب روش‌های مختلف تحلیل مکانی و مجموعه داده‌های بزرگ برای پیش‌بینی روندهای جغرافیایی استفاده شده است. به‌طور خاص، در این تحقیق از مجموعه داده‌های SEDAC برای تحلیل اطلاعات اجتماعی-اقتصادی و از NASA NEOبرای داده‌های آب‌وهوایی استفاده شده است.

فهرست مطالب

1) پیش‌زمینه

1-1)داده‌های مکانی و انواع آن

داده‌های مکانی به دو دسته‌ی اصلی تقسیم می‌شوند :داده‌های برداری و داده‌های شطرنجی .

داده‌های برداری[1]

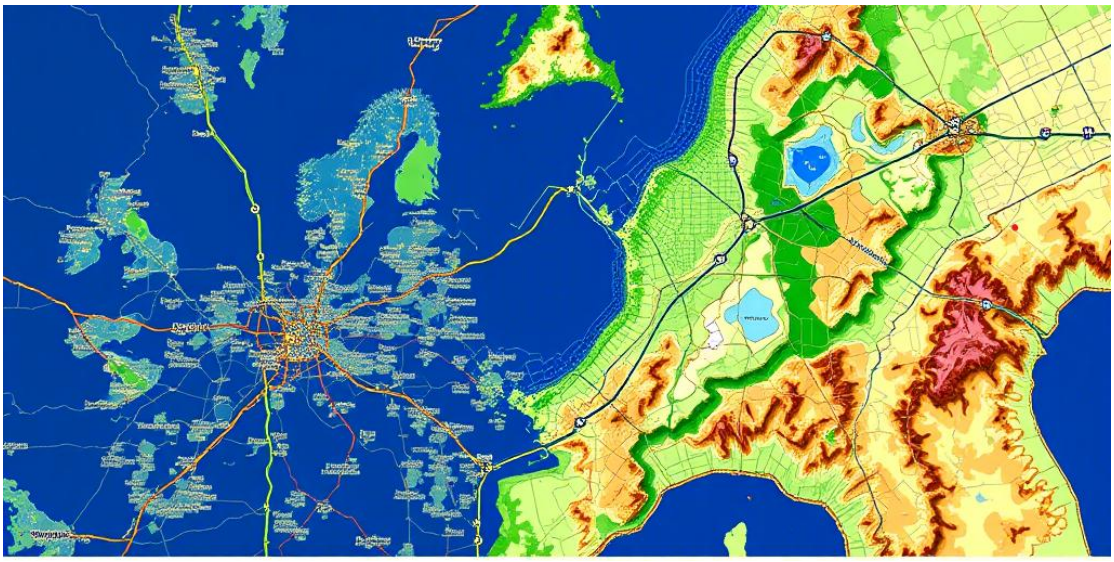
داده‌های برداری از نقاط، خطوط و چندضلعی‌ها تشکیل شده‌اند که نمایانگر مکان های جغرافیایی با مرزهای مشخص هستند. این نوع داده‌ها برای نمایش پدیده‌های گسسته و دارای مرزهای مشخص مانند ساختمان‌ها، جاده‌ها، مرزهای سیاسی و رودخانه‌ها استفاده می‌شوند.

ساختار داده‌های برداری

نقاط : نقاط برای نمایش مکان های که به صورت مجزا و بدون بعد هستند استفاده می‌شوند. برای مثال، موقعیت یک شهر، یک درخت یا یک چاه می‌تواند به صورت یک نقطه نمایش داده شود.

خطوط :خطوط برای نمایش مکتن های خطی مانند جاده‌ها، رودخانه‌ها یا خطوط انتقال برق استفاده می‌شوند.

چندضلعی‌ها : چندضلعی‌ها برای نمایش عوارضی که دارای مساحت هستند مانند ساختمان‌ها، مرزهای سیاسی یا دریاچه‌ها استفاده می‌شوند.



[1] Vector Data

شکل 1. نمونه ای از شکل داده برداری نقشه‌های شهری که شامل خیابان‌ها، ساختمان‌ها و پارک‌ها هستند، نمونه‌ای از داده‌های برداری هستند. هر خیابان به صورت یک خط، هر ساختمان به صورت یک چندضلعی و هر تقاطع به صورت یک نقطه نمایش داده می‌شود.

داده‌های شطرنجی[2]

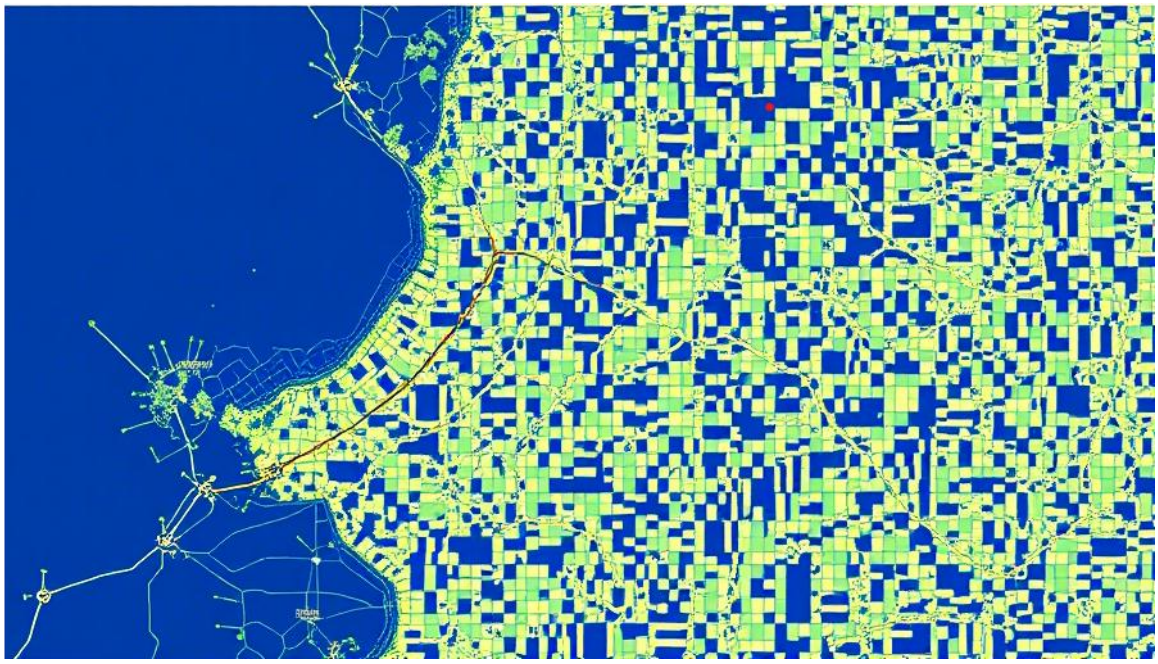
داده‌های شطرنجی را توان به عنوان یک مدل ماتریسی تعریف کرد که داده‌ها را در یک شبکه‌ی منظم از سلول‌ها (پیکسل‌ها) ذخیره می‌کند. هر سلول در این شبکه نمایانگر یک مقدار خاص است که می‌تواند شامل اطلاعاتی مانند ارتفاع، دما، یا میزان بارش باشد. داده‌های شطرنجی برای نمایش پدیده‌های پیوسته و تغییرات تدریجی در سطح زمین مناسب هستند.

ساختار داده‌های شطرنجی

سلول : هر سلول در داده‌های شطرنجی یک مقدار منحصر به فرد دارد که می‌تواند عددی، رنگی یا حتی طبقه‌بندی شده باشد.

رزولوشن : اندازه‌ی هر سلول در داده‌های شطرنجی به عنوان رزولوشن شناخته می‌شود. رزولوشن بالاتر به معنای سلول‌های کوچک‌تر و دقت بیشتر در نمایش داده‌ها است.

باند : داده‌های شطرنجی می‌توانند چندباندی باشند، به این معنی که هر سلول می‌تواند چندین مقدار مختلف را در خود ذخیره کند. برای مثال، تصاویر ماهواره‌ای معمولاً چندباندی هستند و هر باند نمایانگر یک طول موج خاص از نور است.



شکل 2.نمونه از داده شطرنجی، داده‌های شطرنجی معمولاً در سیستم‌های اطلاعات جغرافیایی (GIS) برای نمایش پوشش زمین، تغییرات دمایی و مدل‌های رقومی ارتفاع (DEM) به کار می‌روند.

[2] **Raster Data**

مقایسه داده‌های شطرنجی و برداری

ویژگی	داده‌های برداری	داده‌های شطرنجی
ساختار داده	نقاط، خطوط و چندضلعی‌ها	شبکه‌ی منظم از سلول‌ها (پیکسل‌ها)
نمایش پدیده‌ها	پدیده‌های گسسته (مانند ساختمان‌ها، جاده‌ها)	پدیده‌های پیوسته (مانند ارتفاع، دما)
دقت	دقت بالا، به خصوص در مرزهای مشخص	وابسته به رزولوشن (اندازه‌ی سلول)
حجم داده	حجم داده‌ها معمولاً کوچک‌تر است	حجم داده‌ها معمولاً بزرگ است
کاربردها	نقشه‌برداری شهری، تحلیل شبکه	تحلیل‌های محیطی، تصاویر ماهواره‌ای

مقیاس، وضوح و منطقه‌بندی[3]

مقیاس در نقشه‌ها به عنوان نسبت یک فاصله بر روی نقشه به فاصله واقعی آن در جهان تعریف می‌شود. برای مثال، اگر نقشه‌ای ۱۰۰ متر از واقعیت را در ۱ سانتی‌متر نشان دهد، مقیاس آن ۱:۱۰,۰۰۰ است. وضوح مکانی به حد دقتی که داده‌های مکانی نمایش داده می‌شوند اشاره دارد و در داده‌های وکتوری، میزان جزئیات نمایش داده‌شده در نقاط، خطوط و چندضلعی ها را تعیین می‌کند. برخلاف داده‌های شطرنجی که وضوح آن‌ها بر اساس اندازه پیکسل مشخص است، در داده‌های وکتوری وضوح به سطح دقت اندازه‌گیری و تعداد نقاط استفاده‌شده برای نمایش عوارض بستگی دارد .

منطقه‌بندی (Zonation) یکی از مفاهیم کلیدی در تجزیه و تحلیل مکانی است که برای تقسیم‌بندی داده‌ها به بخش‌های منطقی استفاده می‌شود. در این روش، داده‌های مکانی بر اساس معیارهای مشخص مانند جمعیت، کاربری زمین یا عوامل طبیعی به مناطق مختلف تقسیم می‌شوند. این تقسیم‌بندی می‌تواند تأثیر مستقیمی بر نحوه تحلیل و نتیجه‌گیری داشته باشد. انتخاب نامناسب مرزهای منطقه‌بندی ممکن است منجر به نتایج گمراه‌کننده شود .

تأثیر منطقه‌بندی و تجمیع داده‌ها[4]

برای بررسی تأثیر منطقه‌بندی و تجمیع داده‌ها، جمعیت شاغل و نواحی محل سکونت آن‌ها از مجموعه داده استخراج شده و در مناطق مشخصی تجمیع شدند. برای این کار، شناسه منطقه [5] و شناسه شهر [6] به عنوان مختصات مکانی انتخاب شدند.

[3] Raster Data
[4] Effects of Zonation and Aggregation
[5] DID
[6] TID

در فرآیند تجمیع داده‌ها، در نظر گرفتن میزان دقت داده‌ها و معیارهای منطقه‌بندی اهمیت دارد. انتخاب مقیاس نامناسب یا معیارهای نامناسب برای تجمیع ممکن است باعث ایجاد الگوهای گمراه‌کننده در داده‌ها شود. به عنوان مثال، اگر منطقه‌بندی بر اساس واحدهای اداری انجام شود، ممکن است ارتباطات فضایی واقعی بین مناطق همجوار نادیده گرفته شود. از سوی دیگر، تجمیع داده‌ها ممکن است باعث از بین رفتن اطلاعات جزئی اما مهم شود.

قبل از انجام تحلیل‌ها، حذف داده‌های پرت ضروری است، زیرا این داده‌ها ممکن است باعث تحریف نتایج شوند. برای این کار، روش‌های مختلفی مورد استفاده قرار می‌گیرد که در بخش بعدی مورد بررسی قرار خواهند گرفت. دو روش رایج برای حذف داده‌های پرت، روش انحراف معیار و روش دامنه بین چارکی هستند که در ادامه مورد بحث قرار خواهند گرفت.

1-2) تحلیل داده‌های مکانی

تحلیل داده‌های مکانی شامل روش‌هایی مانند درون‌یابی، خوشه‌بندی، تشخیص الگو و تحلیل شبکه است. این روش‌ها به محققان کمک می‌کنند تا الگوهای پنهان در داده‌ها را کشف کنند و پیش‌بینی‌های دقیق‌تری انجام دهند. برای مثال، درون‌یابی معکوس فاصله و روش همسایه‌ی نزدیک‌ترین از روش‌های پرکاربرد در تحلیل داده‌های مکانی هستند.

چالش‌های تحلیل داده‌های مکانی

با وجود پیشرفت‌های اخیر در حوزه‌ی تحلیل داده‌های مکانی، چالش‌هایی مانند حجم بالای داده‌ها، ناهمگونی داده‌ها و نیاز به روش‌های تحلیلی پیشرفته همچنان وجود دارند. علاوه بر این، داده‌های مکانی مربوط به مناطق خاص مانند خاورمیانه ممکن است با چالش‌های خاصی مانند کمبود داده‌های دقیق و به‌روز مواجه باشند.

1-3) مرور ادبیات

مقیاس، رزولوشن و منطقه‌بندی[7]

مقیاس و رزولوشن دو مفهوم کلیدی در تحلیل داده‌های مکانی هستند. همان‌طور که در اشاره شده است، مقیاس به نسبت فاصله‌ی روی نقشه به فاصله‌ی واقعی در دنیای واقعی اشاره دارد. رزولوشن نیز به اندازه‌ی سلول‌ها در داده‌های شطرنجی اشاره می‌کند. هرچه رزولوشن بالاتر باشد، دقت داده‌ها بیشتر است، اما حجم داده‌ها نیز افزایش می‌یابد.

[7] Scale, Resolution, and Zonation

منطقه‌بندی نیز یک روش کلیدی برای تجمیع داده‌های مکانی است. همان‌طور که در اشاره شده است، منطقه‌بندی می‌تواند تأثیر قابل توجهی بر الگوها و روندهایی که در داده‌ها مشاهده می‌شود، داشته باشد. برای مثال، در تحلیل داده‌های اجتماعی-اقتصادی، منطقه‌بندی می‌تواند به درک بهتر از توزیع جمعیت و درآمد کمک کند.

تشخیص داده‌های پرت

داده‌های پرت می‌توانند تأثیر قابل توجهی بر نتایج تحلیل‌های مکانی داشته باشند. همان‌طور که در اشاره شده است، روش‌های مختلفی برای تشخیص داده‌های پرت وجود دارد، از جمله روش انحراف معیار و روش محدوده‌ی بین چارکی.

ضریب جینی

ضریب جینی یک شاخص کلیدی برای اندازه‌گیری نابرابری درآمدی است. ، این ضریب می‌تواند برای تحلیل توزیع درآمد در مناطق مختلف استفاده شود.

درون‌یابی

درون‌یابی یک روش کلیدی برای پیش‌بینی مقادیر ناشناخته در مناطق بدون داده است. همان‌طور که در [7] اشاره شده است، روش‌های مختلفی برای درون‌یابی وجود دارد، از جمله درون‌یابی معکوس فاصله و روش همسایه‌ی نزدیک‌ترین . در این تحقیق ، از روش درون‌یابی معکوس فاصله برای پیش‌بینی میزان بارش در مناطق مختلف استفاده شده است.

ماتریس فاصله

ماتریس فاصله یک ابزار کلیدی برای تحلیل روابط فضایی بین نقاط مختلف است. ماتریس فاصله می‌تواند برای تحلیل خوشه‌بندی یا ایجاد نقشه‌های حرارتی استفاده شود. در مقاله‌ی از ماتریس فاصله برای تحلیل توزیع جمعیت در مناطق مختلف استفاده شده است.

2)تحلیل داده

تحلیل داده‌های وکتوری استان های کشور های مختلف با استفاده از مفاهیم مقیاس، رزولوشن، و منطقه‌بندی می‌تواند به درک بهتر از توزیع جمعیت، نابرابری درآمدی کمک کند. در مقاله‌ی مرجع ، ازروش‌ها آماری گفته شده برای تحلیل داده‌های سال ۲۰۰۱ استان راجستان هند استفاده شده است، ما با استفاده از داده‌های مشابه تحقیقی مشابه انجام خواهیم داد.

2-1)داده

در قسمت اول این پروژه از داده ای برداری به فرمت shp[8] استفاده می شود. این مجموعه داده ها از مرکز داده ها و کاربردهای اقتصادی- اجتماعی (SEDAC) بازیابی شده است. این مجموعه داده مربوط به ویژگی های اجتماعی-اقتصادی کشور هند ایالت ماهاشترا است. متغییر مورد بررسی ما اشتغال در این ایالت می باشد.

ویژگی	مقدار
تعداد رکورد ها	43,561
تعداد متغییر ها	214
سیستم مختصات (CRS)	+proj=utm +zone=44 +datum=WGS84 +units=m +no_defs
X min	-750000.5
X max	250100.3
Y min	2450000.2
Y max	3305000.8
منبع داده	Socioeconomic Data and Applications Center (SEDAC)
سال داده	2001
محدوده جغرافیایی	ایالت ماهاراشترا، هند

Map of Regions

[8] فرمت (Shapefile) **SHP** یک فرمت رایج برای ذخیره‌سازی داده‌های مکانی برداری در سیستم‌های اطلاعات جغرافیایی (GIS) است که شامل نقاط، خطوط و چندضلعی‌ها برای نمایش عوارض جغرافیایی مانند جاده‌ها، ساختمان‌ها و مرزها می‌شود. این فرمت به دلیل سادگی و سازگاری گسترده، به طور گسترده در نقشه‌برداری و تحلیل‌های مکانی استفاده می‌شود.

3)تشخیص داده های پرت

روش‌های تشخیص داده‌های پرت

داده‌های پرت [9] می‌توانند باعث ایجاد تحلیل‌های نادرست شوند. در این پژوهش، سه روش برای شناسایی داده‌های پرت استفاده شده است:

روش انحراف معیار: شناسایی داده‌هایی که بیش از ۳ انحراف معیار از میانگین فاصله دارند.
روش چارکی :استفاده از دامنه بین چارکی برای حذف داده‌های پرت.

3-1)روش انحراف معیار[10]

برای استفاده از روش انحراف معیار به عنوان یک روش حذف داده‌های پرت، توزیع داده‌ها باید نرمال یا نزدیک به نرمال باشد. داده‌های عددی به دو دسته پیوسته و دسته‌بندی‌شده (گسسته) تقسیم می‌شوند. داده‌های پیوسته مقدارهایی را نشان می‌دهند که روی یک مقیاس اندازه‌گیری می‌شوند، در حالی که داده‌های دسته‌بندی‌شده به مقادیر مشخصی مانند نسبت‌ها یا درصدها اشاره دارند.

توزیع داده‌ها تعیین‌کننده نحوه شناسایی داده‌های پرت است. برای مثال، داده‌های دسته‌بندی‌شده معمولاً دارای توزیع دووجهی هستند و مقادیر خروجی آن‌ها باینری است. اما داده‌های پیوسته معمولاً به شکل یک توزیع نرمال توزیع شده‌اند. در یک توزیع نرمال، مقدار انحراف معیار از میانگین برای توصیف میزان پراکندگی داده‌ها استفاده می‌شود .تابع چگالی احتمال توزیع نرمال در معادله زیر تعریف شده است:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

که در آن:

- σ انحراف معیار است.
- μ میانگین داده‌ها است.

به‌طور کلی، سه انحراف معیار از میانگین به عنوان یک آستانه معمول برای شناسایی و حذف داده‌های پرت در توزیع نرمال استفاده می‌شود. با این حال، در برخی موارد، دو انحراف معیار برای مجموعه داده‌های کوچک‌تر و چهار انحراف معیار برای داده‌های بسیار

[9] Outliers
[10] Standard Deviation Method

بزرگ مناسب‌تر است. به‌طور معمول، داده‌ها ابتدا نرمال سازی می‌شوند تا میانگین آن‌ها صفر و واریانس یک شود. این کار باعث می‌شود که حذف داده‌های پرت بر اساس مقدار Z-score انجام شود.

برای محاسبه میانگین و انحراف معیار جمعیت شاغل، از معادلات زیر استفاده می‌شود:

$$(1) \quad \mu = \frac{1}{n}\sum_{i=1}^{n} w_i$$

$$(2) \quad \sigma = \sqrt{\frac{\sum(w_i-\mu)^2}{n}}$$

که در آن:

- w_i مقادیر مربوط به جمعیت شاغل است.
- n تعداد کل افراد در جمعیت شاغل است.
- انحراف معیار (σ) پراکندگی داده‌ها حول میانگین را اندازه‌گیری می‌کند.

در این مطالعه ، مقدار میانگین و انحراف معیار محاسبه و تقریباً برابر با 879 و 1109 شد. برای تعیین مقادیر پرت، مقدار آستانه به صورت سه برابر انحراف معیار در نظر گرفته می‌شود، زیرا داده‌ها نه بسیار کوچک هستند و نه بیش از حد بزرگ. معادلات زیر گام‌های این فرآیند را نشان می‌دهند:

$$t = k \,.\, \sigma$$

$$l = \mu - t$$

$$u = \mu + t$$

که در آن:

- t مقدار آستانه برای حذف داده‌های پرت است.
- l حد پایین داده‌های معتبر است.
- u حد بالای داده‌های معتبر است.
- $k = 3$ به عنوان مقدار پیش‌فرض در نظر گرفته شده است.

میانگین:(μ)	انحراف معیار:(σ)	ضریب:(k)
879.213≈879	1108.974≈1109	3
حد بالا:(u)	مقدار آستانه:(t)	حد پایین:(l)
3690522	3327	-3688764

بنابراین، هر داده کمتر از l یا بیشتر از u به عنوان داده پرت شناخته می شود.

2-3)روش دامنه بین چارکی[11]

یکی از روش‌های آماری رایج برای شناسایی داده‌های پرت، روش دامنه بین چارکی است. برخلاف روش انحراف معیار که برای داده‌های با توزیع نرمال مناسب است، روش دامنه بین چارکی برای داده‌هایی که لزوماً توزیع نرمال ندارند، کارآمدتر می‌باشد. این روش بر اساس محدوده‌ای که بین صدک ۲۵م و صدک ۷۵م قرار دارد، مقدارهای پرت را شناسایی می‌کند.

محاسبه دامنه بین چارکی

دامنه بین چارکی با استفاده از رابطه زیر محاسبه می‌شود :

- صدک ۲۵م که مقدار داده‌ای را نشان می‌دهد که **٪۲۵ داده‌ها کمتر از آن و ٪۷۵ داده‌ها بیشتر از آن هستند.**:
- صدک ۷۵م که مقدار داده‌ای را نشان می‌دهد که **٪۷۵ داده‌ها کمتر از آن و ٪۲۵ داده‌ها بیشتر از آن هستند.**:

صدک	٪0	٪25	٪50	٪75	٪100
مقدار	0.0	218	441	798	1743041

تعیین آستانه برای داده‌های پرت

با استفاده از مقداردامنه بین چارکی، محدوده‌ای برای شناسایی داده‌های پرت تعیین می‌شود. برای این منظور، مقداردامنه بین چارکی در یک ضریب ضرب شده و از صدک‌های ۲۵م و ۷۵م کم یا به آن اضافه می‌شود:

$$t = k.lQR$$

$$l = q25 - t$$

$$u = q75 + t$$

که در آن:

- t مقدار آستانه یا مقدار برش است.
- l حد پایین داده‌های معتبر.

[11] Interquartile Range - IQR

- u حد بالای داده‌های معتبر.
- $k = 1.5$ مقدار به‌عنوان مقدار استاندارد پذیرفته‌شده است.

در این مطالعه، ابتدا صدک‌های 25م و 75م برای داده‌های جمعیت شاغل محاسبه شدند. سپس با استفاده از مقدار دامنه بین چارکی برای تعیین حدود داده‌های پرت به کار گرفته شد. داده‌هایی که کمتر از یا بیشتر از بودند، به‌عنوان داده‌های پرت حذف شدند. جدول - مقادیر محاسبه‌شده را نمایش می‌دهد.

q25	q75	IQR	t
218	798	510	1530
	l	u	
	-1315	2328	

اگر داده ای مجود داشت که کمتر l از یا بیشتر از u بودند، به‌عنوان داده‌های پرت حذف شدند.

روش دامنه بین چارکی روشی کارآمد برای شناسایی داده‌های پرت در مجموعه داده‌هایی است که توزیع نرمال ندارند. این روش به‌ویژه در مواقعی که داده‌ها دارای چولگی باشند، بسیار مفید است. با استفاده از این روش، داده‌های پرت از مجموعه داده حذف شده و تحلیل‌های مکانی با دقت بیشتری انجام می‌شود .

3-3)نتیجه‌گیری در مورد داده‌های پرت و پیش بینی آنها

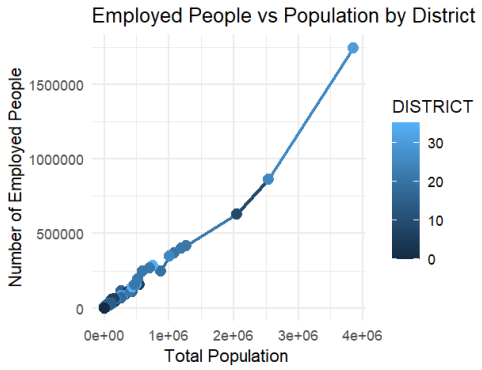
هر دو روش انحراف معیار و روش محدوده بین‌چهارکی می‌توانند برای داده‌های چندمتغیره با تعیین محدوده‌ها برای هر متغیر در مجموعه داده استفاده شوند و مقادیری که خارج از این محدوده‌ها قرار می‌گیرند را به‌عنوان داده‌های پرت حذف کنند. هر دو روش انحراف معیار و محدوده بین‌چهارکی محدوده‌هایی را ارائه دادند که به حذف داده‌های پرت کمک کردند، اما به نظر می‌رسد روش محدوده بین‌چهارکی از نظر آماری دقیق‌تر از روش انحراف معیار است. برای بررسی داده‌ها در هر دو بعد مکانی و غیرمکانی، نمودارهای در بخش بعد ترسیم شده اند که به تحلیل بهتر برای شاخص اقتصادی مانند اشتغال کمک می کنند.

4-3) پردازش داده های ایالت ماهاشترا

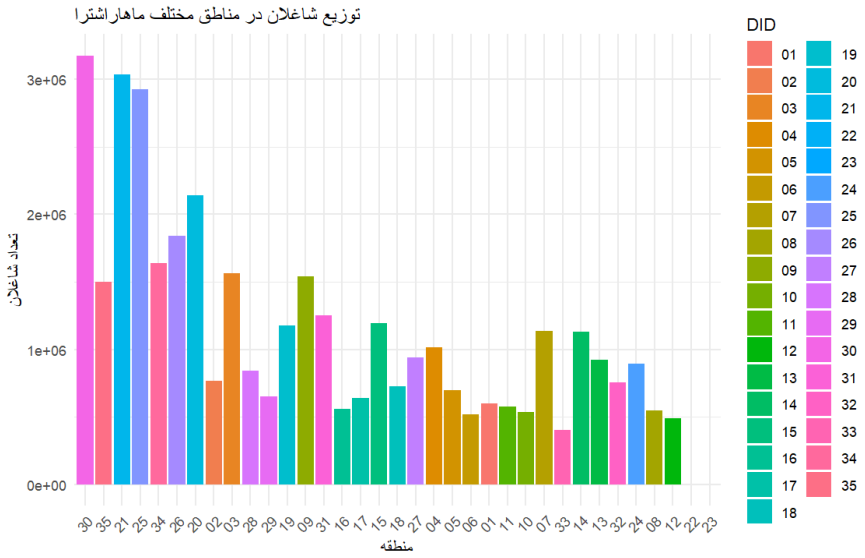
با رسم نمودار های جمعیت و اشتغال و توزیع اشغال بر اساس منطقه می توان نتیجه های گرفت. نمودار ها ، نشان می‌دهند که الگوی مکانی مشخصی در توزیع اشتغال مناطق وجود دارد وجود دارد. بر اساس نمودار اول به اشتغال بر اساس جمعیت را نشان

می دهد می توان نتیجه گرفت بر اساس جمعیت مناطق این استان با روندی خطی با افزایش جمعیت اشتغال افزایش می یابد.

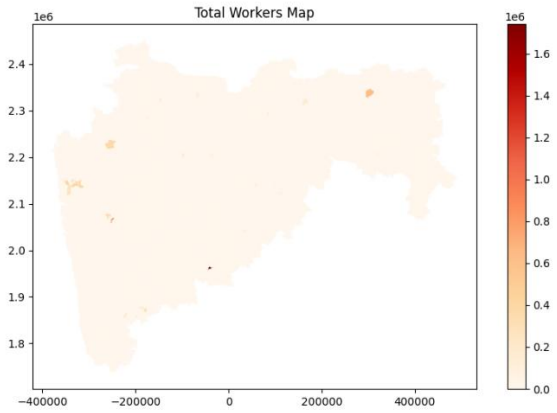
(یکسان نبودن جمعیت مناطق)توزیع افراد شاغل در مناطق نشان می دهد توزیع کار و افراد شاغل بسیار نابرابر است. و بیشترین تعداد افراد شاغل در مناطق 30و 21 و25 قرار دارد. در نهایت توزیع شاغلین در نقشه نشان می هد می توان دید اشتغال در این ایالت نسبت به جمعیت و مساحت بسیار کم است



نمودار1.نمودار خطی که تعداد جمعیت شاغل را نشان می دهد



نمودار2. توزیع افراد شاغل در مناطق مختلف ایالت



نمودار3. توزیع افراد شاغل در ایالت

3-5)شاخص جینی(Gini Coefficient)

ضریب جینی یکی از معیارهای مهم آماری است که برای اندازه‌گیری نابرابری در توزیع درآمد یا ثروت درون یک کشور یا منطقه به کار می‌رود. این شاخص برای اندازه‌گیری پراکندگی است. شاخص جینی نابرابری بین مقادیر یک توزیع فراوانی را کمّی می‌کند که در این مورد، نابرابری درآمد است. علاوه بر این، شاخص جینی 0 به معنای برابری کامل است، جایی که همه متغیرها برابر هستند. در مقابل، شاخص جینی 1 نشان‌هنده بیشترین نابرابری بین متغیرها است .

ضریب جینی = 0 :نشان‌دهنده برابری کامل است، یعنی تمام خانوارها یا افراد درآمد یا ثروت برابری دارند.

ضریب جینی = 1 : نشان‌دهنده حداکثر نابرابری است، یعنی یک خانوار یا فرد تمام درآمد یا ثروت را در اختیار دارد.

به‌عنوان مثال، یکی از اهداف سیاست‌های کمونیستی کاهش شاخص جینی به 0 است.

برای محاسبه شاخص جینی از معادله زیر استفاده می شود:

$$G = \frac{2\sum_{i=1}^{n} i\, y_i}{n\Sigma_{i=1}^{n} y_{i=1}} - \frac{n+1}{n}$$

در این مقاله، با فرض اینکه تعداد کارکنان در هر خانوار مرتبط با سطح درآمد آن خانوار است، از ضریب جینی برای محاسبه نابرابری درآمدی در ایالت ماهاراشترا در سال 2001 استفاده شده است. که برابر 0.6419071532334081 شد که نشان از نابرابرای توزیع درآمد در این ایالت هند است و نتیجه گیری ما را ثابت می کند. نتایج نشان داد که نابرابری درآمدی در این ایالت نسبتاً بالا است و نیاز به سیاست‌های اقتصادی برای کاهش این نابرابری وجود دارد.

3-6)فاصله، تأثیر مکانی، و همبستگی

فاصله یک معیار است که نشان می‌دهد چقدر اشیاء از یکدیگر دور هستند. یکی از روش‌های متداول برای توصیف فاصله، استفاده از **ماتریس فاصله** [12]است. ماتریس فاصله شامل مقادیر عددی است که فاصله بین تمام اشیاء را نشان می‌دهد.

در مورد داده‌های ما:

مقادیر فاصله به صورت درجه‌های مختصات جغرافیایی [13] ثبت شده‌اند، بنابراین نیاز به تبدیل آنها به واحد مناسب‌تر (مثل کیلومتر) وجود دارد.

اولین گام، کاهش اندازه داده‌های مختصاتی بود زیرا مجموعه داده اصلی شامل 40,000 ورودی است که باعث ایجاد یک ماتریس فاصله با حجم حدود 11.9 گیگابایت می‌شود، که برای تحلیل قابل مدیریت نیست.

پیش‌پردازش داده‌ها

برای کاهش حجم داده‌ها:

1. فیلتر کردن داده‌ها بر اساس شناسه ناحیه:[14]
   مجموعه داده به 1570 ورودی کاهش یافت که فقط شامل ورودی‌هایی با شناسه ناحیه برابر با 26 است.
2. ** down-sampling:**
   سپس، مجموعه داده 1570 ورودی‌ای به 100 ورودی کاهش یافت، که هر ورودی به صورت تصادفی انتخاب شده است.
3. ساخت ماتریس فاصله:
   پس از انتخاب ورودی‌ها، داده‌ها در یک ماتریس فاصله ترکیب شدند و این ماتریس به یک ماتریس عادی تبدیل شد.

	1	2	3	4	5
1	0.000	3609.361	16316.869	18541.482	11493.070
2	3609.361	0.000	12845.461	14981.418	7979.095
3	16316.869	12845.461	0.000	2788.382	4880.120
4	18541.482	14981.418	2788.382	0.000	7075.914
5	11493.070	7979.095	4880.120	7075.914	0.000

ماتریس فاصله در 5 نمونه از فرم نهایی داده ها

کاربردهای ماتریس فاصله

ماتریس فاصله در سیستم‌های غیرجغرافیایی نیز به طور گسترده استفاده می‌شود.

[12]Distance Matrix
[13] degrees of coordinates
[14] District ID

سیستم‌های غیرجغرافیایی به‌طور گسترده از ماتریس‌های فاصله استفاده می‌کنند. به‌ویژه، این ماتریس معمولاً برای تولید نمودار خوشه‌بندی [15] یا دندروگرام‌ها[۱۶] استفاده می‌شود.

ایجاد یک معیار برای تأثیر مکانی بین موجودیت‌ها، یک فرآیند کلیدی در آمار مکانی است. این معیار معمولاً به‌عنوان ماتریس وزن‌های مکانی توصیف می‌شود و به‌عنوان تابعی از همسایگی تفسیر می‌شود.

تأثیر مکانی بسیار پیچیده است و هیچ روش قطعی برای کمّی‌سازی آن وجود ندارد. با این حال، مانند بسیاری از معیارها در آمار، می‌توان آن را به روش‌های مختلف ارزیابی کرد. ایده همبستگی مکانی در آمار مکانی بسیار مهم است. این مفهوم هم مبهم است (زیرا تحلیل آزمون‌های آماری را دشوار می‌کند) و هم مفید است (زیرا امکان درونیابی مکانی را فراهم می‌کند).

4)درون یابی[۱۷]

درون‌یابی یک روش است که به ما اجازه می‌دهد مقادیر نامعلوم داده‌ها را با استفاده از مقادیر شناخته‌شده پیش‌بینی کنیم. در سیستم اطلاعات جغرافیایی، بسیاری از روش‌های درون‌یابی مورد استفاده قرار می‌گیرند، از جمله درون‌یابی خطی[18].

همبستگی فضایی [19] برای هر متغیر معنی‌دار وجود دارد و می‌تواند در تحلیل‌های آماری مانعی باشد. با این حال، مزایای این همبستگی زمانی ظاهر می‌شود که سعی داریم مقادیر را در مناطقی که درباره آنها داده قبلی وجود ندارد، پیش‌بینی کنیم.

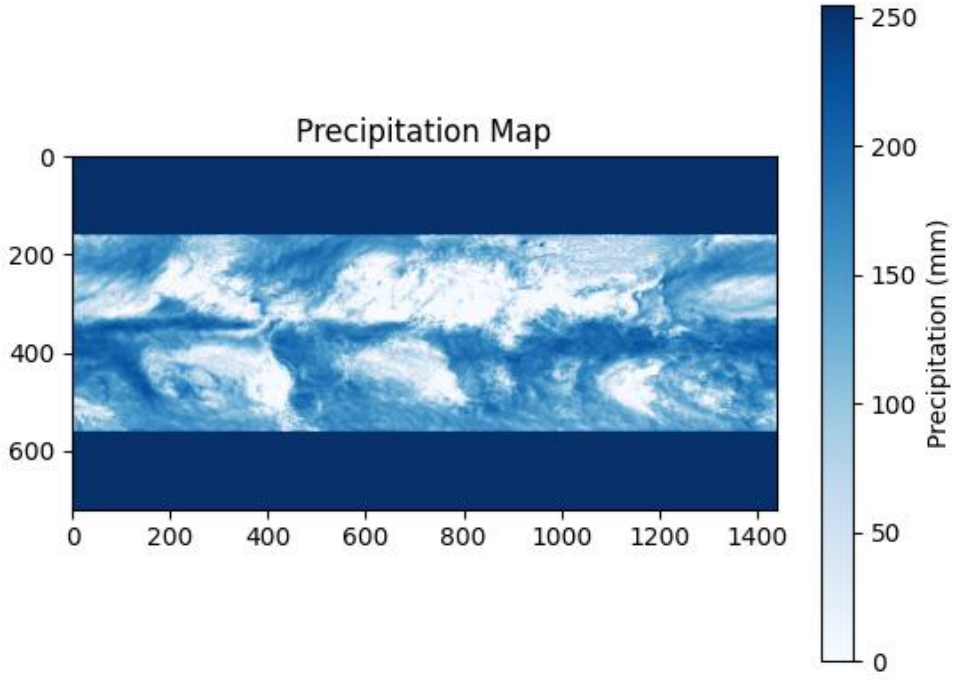
داده‌های مورد استفاده: برای این بخش از تحقیق، از داده‌های ناسا[20] استفاده شده است. این داده‌ها شامل مجموعه‌ای از اطلاعات برای جهان در سال 2016 است. به طور خاص، از داده‌های تاریخی مربوط به نرخ بارندگی یک ماه در سطح جهان استفاده شده است.

شکل 8 نمایشی از داده‌های شطرنجی است که توزیع بارندگی روی سطح وادی مرگ را نشان می‌دهد. این شکل یک مثال از نحوه استفاده از بردارهای عددی برای توصیف مکان‌ها و توسعه نقشه‌های اولیه است.

---

[15] Cluster Diagrams
[16] Dendrograms
[17] Interpolation
[18] Linear Interpolation
[19] Spatial Correlation

[20] https://neo.gsfc.nasa.gov



شکل8.[۲۱]توزیع بارش در February

4-1)مدل نال (Null Model) و ریشه میانگین مربع خطا(RMSE)

مدل نال [22] یک مدل آماری است که با استفاده از نمونه‌های تصادفی از یک توزیع مشخص ساخته شده و شامل برخی ویژگی‌های ثابت و برخی ویژگی‌های تصادفی است. مدل نال برای تعیین توزیع آماری داده‌ها یا تصادفی‌سازی مشاهدات به کار می‌رود و هدف آن پیش‌بینی نتیجه یک تابع تصادفی بدون تعیین تمام متغیرهای آن است. این مدل در تحلیل داده‌های مکانی و همچنین برای ایجاد یک خط پایه جهت مقایسه با سایر مدل‌های پیچیده‌تر مورد استفاده قرار می‌گیرد. در واقع، مدل نال فرض می‌کند که هیچ رابطه آماری معناداری بین متغیرهای مشاهده‌شده وجود ندارد، مگر آنکه شواهد آماری خلاف آن را اثبات کنند.

[21] نواحی رنگی روی نقشه های بالا نشان می دهد که کجا و چه مقدار بارندگی در سراسر جهان بر حسب میلی متر باریده است. حدود دو سوم کل بارندگی در امتداد خط استوا یا نزدیک آن می بارد و بیشتر از خشکی روی اقیانوس می بارد. این نکته کلیدی اندازه گیری دقیق بارندگی در مقیاس جهانی را تا همین اواخر برای دانشمندان دشوار می کرد.

[22] Null Model

از مدل نال می‌توان برای مقایسه الگوهای مکانی واقعی با یک الگوی تصادفی استفاده کرد. این مقایسه به محققان اجازه می‌دهد که الگوهای مشاهده‌شده را از الگوهای تصادفی متمایز کنند و بفهمند که آیا توزیع مکانی داده‌ها از یک فرآیند تصادفی پیروی می‌کند یا تحت تأثیر عواملی خاص قرار دارد.

برای ارزیابی دقت مدل نال، از ریشه میانگین مربع خطا [۲۳]استفاده شده است . مقدار خطای مدل را با محاسبه انحراف مربع داده‌های مشاهده‌شده از مقادیر پیش‌بینی‌شده تعیین می‌کند. این معیار به‌عنوان یکی از رایج‌ترین روش‌های سنجش دقت مدل‌های پیش‌بینی مورد استفاده قرار می‌گیرد.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^2}$$

که در آن:

- O_i مقادیر مشاهده‌شده
- S_i مقادیر پیش‌بینی‌شده
- n تعداد کل مشاهدات

RMSE یک معیار مناسب برای اندازه‌گیری دقت مدل‌های پیش‌بینی است، اما به دلیل اینکه به مقیاس داده‌ها وابسته است، نمی‌توان از آن برای مقایسه متغیرهای مختلف استفاده کرد .به عبارت دیگر، RMSEفقط برای مقایسه خطای مدل‌های مختلف که یک متغیر مشخص را پیش‌بینی می‌کنند، مناسب است.

4-2)فاصله معکوس[24]

یکی از تفاوت‌های برجسته بین آمار فضایی و آمار عددی، ادغام مستقیم داده‌های فضایی و روابط آنها در محاسبات ریاضی است. فاصله معکوس[25] ، زمان سفر[26] ، فاصله ثابت، نزدیکترین همسایه[27] ، و پیوستگی از جمله مفاهیم متداول برای توصیف روابط فضایی هستند . روش‌های تعریف این مفاهیم به نوع اندازه‌گیری بستگی دارند. فاصله معکوس به‌طور معمول برای ارزیابی خوشه‌بندی مناسب‌تر است. مدل مفهومی تعاملات مکانی با گزینه‌های فاصله معکوس می‌تواند شامل موانع یا کاهش فاصله باشد. در این چارچوب، هر ویژگی تأثیراتی بر سایر ویژگی‌ها دارد، اما با کاهش متغیرها، تأثیر آن‌ها کمتر می‌شود.

[23] Root Mean Square Error - RMSE
[24] Inverse Distance
[25] Inverse Distance
[26] Journey Time
[27] K Nearest Neighbors

فاصله معکوس یکی از بهترین روش‌ها برای ارزیابی خوشه‌بندی[28] است. در این روش، هر عنصر بر تمام عناصر دیگر تأثیر می‌گذارد، اما با افزایش فاصله، میزان تأثیر کاهش می‌یابد. برای کاهش تعداد محاسبات لازم در داده‌های بزرگ ، مقدار "باند فاصله[29]" یا "حداکثر فاصله[30] "محاسبه می‌شود.[17] طبق تعریف در ویکی‌پدیا، فرمول فاصله بین دو نقطه مختصاتی در معادله زیر آمده است.

$$l = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

$(y_i - y_j)$ و $(x_i - x_j)$ مختصات نقاط i و j هستند.

روش وزن‌دهی فاصله معکوس [31] یکی از روش‌های پیش‌بینی مقادیر نامعلوم است که از نزدیکی نقاط به یکدیگر برای تعیین وزن استفاده می‌کند. این روش متغیر هدف را با اختصاص وزن بیشتر به نقاطی که نزدیک‌تر هستند، ارزیابی می‌کند. در این روش به داده‌های قبلی برای پیش‌بینی فضایی نیازی نیست. با این حال، باوجود راحتی آن، روش‌های برآورد فاصله معکوس به‌شدت به نوع پایگاه داده، تعداد گره‌های استفاده‌شده در برآورد و توان فاصله اعمال‌شده در وزن‌دهی وابسته است. علاوه بر این، استفاده از درونیابی فاصله معکوس در مقایسه با دیگر کاربردهای عملی ممکن است برای برآوردهای معتبرتر از ساختار فضایی داده‌ها مناسب‌تر باشد.

V2	V3	V4	V5	V6	V7
1e+10	1e+10	1e+10	1e+10	0.0001399191	0.0001402891
Inf	1e+10	1e+10	1e+10	0.0001399191	0.0001402891
1e+10	Inf	1e+10	1e+10	0.0001399191	0.0001402891
1e+10	1e+10	Inf	1e+10	0.0001399191	0.0001402891

جدول1 ماتریس معکوس فاصله

ماتریس فاصله معکوس [32] شامل مقادیر معکوس فاصله بین تمام جفت‌های نقاط است. برخی مقادیر Inf(بی‌نهایت) وجود دارند که ناشی از تقسیم بر صفر است (فاصله خود یک نقطه با خودش). این مقادیر در حالت کلی با "NA" جایگزین می‌شوند. هرچه نقاط به یک نزدیکتر باشند فاصله کوتاه تر و هرچه به صفر نزدیکتر باشند فاصله بزرگتز است. فواصل در این نمونه از یک تا 5 بسیار بزرگ و از 6 تا7 بسیار کوچک اند. هرچند عدد 1e+10 عددی نا معقول در این ماتریس است و نشان از کیفیت پایید داده ورودی دارد.

[28] Clustering
[29] Distance Band
[30] Threshold Distance
[31] Inverse Distance Weighting - IDW
[32] Inverse Distance Matrix

ماتریس‌های وزن فضایی معمولاً برای اینکه مجموع سطرها برابر شود، نرمال می‌شوند. بنابراین، گام بعدی این است که سطرها را بر مجموع آن‌ها تقسیم کنیم. در نهایت، پس از تمیزکاری و نرمال‌سازی، ماتریس فاصله معکوس مشابه جدول2 خواهد بود. برای اعتبارسنجی نتایج، می‌توانیم مجموع سطرها را حساب کنیم و از نرمال‌سازی داده‌ها اطمینان حاصل کنیم.

V1	V2	V3	V4	V5	V6	V7
NA	1	1	1	1	1.399191e-14	1.402891e-14
1	NA	1	1	1	1.399191e-14	1.402891e-14
1	1	NA	1	1	1.399191e-14	1.402891e-14
1	1	1	NA	1	1.399191e-14	1.402891e-14

جدول 2 ماریس فاصله معکوس نرمال شده

در نهایت، درونیابی فاصله معکوس یکی از ساده‌ترین و پرکاربردترین تکنیک‌های درونیابی فضایی قطعی است. درGIS ، وزن‌دهی فاصله معکوس یکی از روش‌های پایه‌ای و گسترده‌استفاده‌شده برای درون‌یابی فضایی است. این روش ترکیبی از مفاهیم نزدیکی و تغییرات تدریجی روی سطح را ارائه می‌دهد. در حالت کلی، IDW میانگین وزن‌دار داده‌های نمونه‌ای در یک منطقه جستجو[۳۳]است.[17, 20] فرمول محاسبه وزن‌دهی فاصله معکوس به‌صورت زیر است:

$$X = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

X مقدار مکان ناشناخته.

w_i وزن.

x_i مقدار نقطه شناخته‌شده.

3-4) الگوریتم نزدیک‌ترین همسایگان[34]

الگوریتم نزدیکترین همسایگان یک روش ساده و قدرتمند است که برای طبقه‌بندی یا پیش‌بینی داده‌ها استفاده می‌شود. این الگوریتم به جای آموزش یک مدل پیچیده، از داده‌های موجود برای تصمیم‌گیری استفاده می‌کند. به همین دلیل، برای داده‌های بزرگ و جریان‌های داده‌ای (مثل داده‌های زنده) بسیار مناسب است.

برخلاف سایر تکنیک‌های رایج، الگوریتم نزدیک‌ترین همسایگان نیازی به پیش‌محاسبه طبقه‌بندی ندارد که آن را برای جریان‌های داده مناسب می‌سازد . زمانی که کارایی و دقت مهم‌ترین عوامل هستند، KNN گزینه‌ای عالی است. علاوه بر این، KNN نسبت به سایر روش‌ها مزیت دارد زمانی که یک طبقه‌بند عمر عملکردی کافی نداشته باشد، مانند حالت جریان‌های داده که سرعت ورود داده‌ها بالاست و مدل باید به طور مداوم در حال آموزش باشد.

[33] Search Zone

[34] K-Nearest Neighbors - KNN

بیشتر داده‌های مکانی در قالب فایل "Band Sequential[35]" (BSQ) هستند [21]. معیارهای آماری داخل یک باند به صورت "رشته‌ای" گروه‌بندی می‌شوند که از ناحیه مکانی موجود در مجموعه داده پیروی می‌کنند. برای مجموعه داده ما، مراحل زیر مورد نیاز است.

ابتدا باید معیار فاصله مناسب تعیین شود. سپس k نزدیک‌ترین همسایه با استفاده از معیار انتخاب‌شده پیدا شوند. در مرحله بعد، کلاس اکثریت k-نزدیک‌ترین همسایه محاسبه می‌شود. در نهایت، باید کلاس‌های پیدا شده به داده‌هایی که قرار است طبقه‌بندی شوند، اختصاص داده شوند. برای این ماتریس مقدار k، 2 در نظر گرفته شده است.

knn_result	23	37	46	54	60	65	77	93	111	113	119	126	127	131	132	136	137	139	140	142	147	148	149	150	151	152	155	157	165
23	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
46	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	1	0	0	0	0
60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
73	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
77	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
83	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
93	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
98	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
102	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
105	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
106	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
109	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
110	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
113	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

ماتریس KNN

دقت این مدل برای داده های ما 0.032258064516129 شده است. دقت پایین KNN نشان‌دهنده چالش‌هایی در انتخاب ویژگی‌ها، تعداد همسایگان، یا کیفیت داده‌های ورودی است.

با توجه به تحلیل‌های انجام‌شده در این پروژه، می‌توان نتیجه گرفت که تحلیل داده‌های مکانی نقشی کلیدی در درک الگوهای جغرافیایی، اجتماعی و اقتصادی دارد. در این تحقیق، با استفاده از دو منبع داده‌ی مختلف، شامل SEDAC برای تحلیل اطلاعات اجتماعی-اقتصادی و NASA NEO برای داده‌های آب‌وهوایی، سعی شد تا روش‌های مختلف تحلیل داده‌های مکانی بررسی و بهبود یابد.

تحلیل داده‌های مکانی بزرگ در سیستم‌های اطلاعات جغرافیایی (GIS) به عنوان یک ابزار قدرتمند برای درک الگوهای مکانی و تصمیم‌گیری‌های استراتژیک در حوزه‌های مختلف از جمله برنامه‌ریزی شهری، مدیریت منابع طبیعی و تحلیل‌های اجتماعی-اقتصادی شناخته می‌شود. این تحقیق نشان داد که با استفاده از روش‌های تحلیلی پیشرفته، می‌توان از داده‌های مکانی برای

[35] یک فرمت ذخیره‌سازی داده‌های رستری است که در آن مقادیر مربوط به هر باند (کانال) به‌صورت متوالی و جداگانه ذخیره می‌شوند. این مدل به‌گونه‌ای سازماندهی شده که ابتدا تمام داده‌های باند اول، سپس باند دوم و به همین ترتیب تا آخرین باند قرار می‌گیرند. این روش به دلیل سادگی و دسترسی سریع به اطلاعات هر باند، در پردازش تصاویر چندطیفی مانند داده‌های سنجش از دور کاربرد فراوانی دارد.

پیش‌بینی و تصمیم‌گیری‌های دقیق‌تر استفاده کرد. با این حال، برای دستیابی به نتایج دقیق‌تر و کاربردی‌تر، نیاز به بهبود کیفیت داده‌ها و استفاده از روش‌های تحلیلی پیشرفته‌تر وجود دارد.

مقاله مرجع: Analysis of Spatial Big Data for Geographical Information Systems[36]

[36] https://www.researchgate.net/publication/360058688_Analysis_of_Spatial_Big_Data_for_Geographical_Information_Systems

کد های R مورد استفاده در این مقاله

کد های بخش3 تشخیص داده های پرت

```
library(sf)
library(dplyr)

shp_file <- "مسیر فایل"
data <- st_read(shp_file)
print(head(data))

variable <- data$TOT_WORK_P

mean_value <- mean(variable, na.rm = TRUE)
sd_value <- sd(variable, na.rm = TRUE)

threshold_high <- mean_value + (3 * sd_value)
threshold_low <- mean_value - (3 * sd_value)

outliers_sd <- data %>% filter(variable < threshold_low | variable > threshold_high)
st_write(outliers_sd, "outliers_sd.shp", delete_layer = TRUE)

Q1 <- quantile(variable, 0.25, na.rm = TRUE)
Q3 <- quantile(variable, 0.75, na.rm = TRUE)
IQR_value <- Q3 - Q1

threshold_high <- Q3 + (3 * IQR_value)
threshold_low <- Q1 - (3 * IQR_value)
```

```
outliers_iqr <- data %>% filter(variable < threshold_low | variable > threshold_high)
```

کد بخش 3 شاخص جینی کد پایتون

```
import geopandas as gpd
import matplotlib.pyplot as plt
import pysal.lib as ps
from pysal.explore import esda
from pysal.viz import splot

shapefile_path = "مسیر فایل"
data = gpd.read_file(shapefile_path)

def gini_coefficient(x):
    x = np.array(x)
    x = x[~np.isnan(x)]
    mad = np.abs(np.subtract.outer(x, x)).mean()
    rmad = mad / np.mean(x)
    gini = 0.5 * rmad
    return gini

gini = gini_coefficient(data["TOT_WORK_P"])
print(f"Gini Coefficient for Total Workers: {gini}")
```

کد R بخش 3 ماتریس فاصله

```
library(sp)
library(dplyr)
library(tmap)
library(sf)
shp_data$centroid <- st_centroid(shp_data$geometry)
coords <- st_coordinates(shp_data$centroid)
distance_matrix <- as.matrix(dist(coords))
print(distance_matrix[1:5, 1:5])
```

کد پایتون بخش 4 بدست آوردن توزیع بارش (نقشه بارش)

```
import rasterio
import numpy as np
import matplotlib.pyplot as plt

with rasterio.open("مسیر فایل") as src:
    precipitation = src.read(1)
    transform = src.transform
    crs = src.crs

x = np.arange(transform[2], transform[2] + transform[0] * precipitation.shape[1],
transform[0])
y = np.arange(transform[5], transform[5] + transform[4] * precipitation.shape[0],
transform[4])
xx, yy = np.meshgrid(x, y)

points = np.column_stack((xx.flatten(), yy.flatten()))
```

```
values = precipitation.flatten()


mask = ~np.isnan(values)
points = points[mask]
values = values[mask]
#نقشه بارش
import matplotlib.pyplot as plt
plt.imshow(precipitation, cmap="Blues")
plt.colorbar(label="Precipitation (mm)")
plt.title("Precipitation Map")
plt.show()
```

کد های بخش 4 ماتریس معکوس فاصله و ماتریس معکوس فاصله نرمال شده

```
library(raster)
library(class)

raster_data <- raster("مسیر فایل")


data_matrix <- as.matrix(raster_data)
data_matrix <- na.omit(data_matrix)


distance_matrix <- as.matrix(dist(data_matrix))


inverse_distance_matrix <- 1 / (distance_matrix + 1e-10)


normalized_inverse_distance_matrix <- inverse_distance_matrix /
max(inverse_distance_matrix)
```

```
head(inverse_distance_matrix, n = 1)
head(normalized_inverse_distance_matrix, n = 5)
```

ماتریس KNN

```
library(raster)
library(class)
raster_data <- raster("مسیر فایل")
data_matrix <- as.matrix(raster_data)
data_matrix <- na.omit(data_matrix
labels <- as.vector(raster("مسیر فایل"))
labels <- na.omit(labels


print(nrow(data_matrix)
print(length(labels))
set.seed(123)
sampled_indices <- sample(1:length(labels), nrow(data_matrix))
sampled_labels <- labels[sampled_indices]
set.seed(123)
train_indices <- sample(1:nrow(data_matrix), 0.7 * nrow(data_matrix))
train_data <- data_matrix[train_indices, ]
test_data <- data_matrix[-train_indices, ]
train_labels <- sampled_labels[train_indices]
test_labels <- sampled_labels[-train_indices]
k <- 10
knn_result <- knn(train = train_data, test = test_data, cl = train_labels, k = k)
accuracy <- sum(knn_result == test_labels) / length(test_labels)
confusion_matrix <- table(knn_result, test_labels)
```

```
print(confusion_matrix)
```