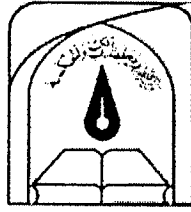


٢١٥١٣



دانشگاه تربیت مدرس

دانشکده علوم پایه

پایان نامه دوره کارشناسی ارشد آمار

روشهای شناسایی داده‌های دورافتاده فضایی

۵ / ۵۱ / ۱۳۸۱

توسط

علی محمدیان مصمم

استاد راهنما

دکتر محسن محمدزاده

وزارتخانه‌های آموزش عالی و
تربیت مدرس

بهمن ماه ۱۳۸۰





۳۱/۵/۸۰

تأییدیه اعضای هیأت داوران حاضر در جلسه دفاع از پایان نامه کارشناسی ارشد

اعضای هیئت داوران نسخه نهایی پایان نامه خانم/ آقای علی محمدیان مصمم

تحت عنوان: روشهای شناسایی داده‌های دور افتاده فضایی

را از نظر فرم و محتوا بررسی نموده و آنرا برای اخذ درجه کارشناسی ارشد مورد تایید قرار دادند.

امضاء	رتبه علمی	نام و نام خانوادگی	اعضای هیأت داوران
	استادیار	آقای دکتر محسن محمدزاده	۱- استاد راهنما
-	-	-	۲- استاد مشاور
	استادیار	آقای دکتر عباس گرامی	۳- استاد ناظر
	دانشیار	آقای دکتر عین‌الله پاشا	۴- استاد ناظر
	دانشیار	آقای دکتر عین‌الله پاشا	۵- نماینده تحصیلات تکمیلی



بسمه تعالی

آیین نامه چاپ پایان نامه (رساله) های دانشجویان دانشگاه تربیت مدرس

نظر به اینکه چاپ و انتشار پایان نامه (رساله) های تحصیلی دانشجویان دانشگاه تربیت مدرس، مبین بخشی از فعالیتهای علمی - پژوهشی دانشگاه است بنابراین به منظور آگاهی و رعایت حقوق دانشگاه، دانش آموختگان این دانشگاه نسبت به رعایت موارد ذیل متعهد می شوند:

ماده ۱ در صورت اقدام به چاپ پایان نامه (رساله) ی خود، مراتب را قبلاً به طور کتبی به «دفتر نشر آثار علمی» دانشگاه اطلاع دهد.

ماده ۲ در صفحه سوم کتاب (پس از برگ شناسنامه)، عبارت ذیل را چاپ کند:

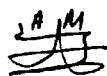
کتاب حاضر، حاصل پایان نامه کارشناسی ارشد / رساله دکتری نگارنده در رشته ^{کار} _{است} که در سال ۱۳۸۰ در دانشکده علوم پایه دانشگاه تربیت مدرس به راهنمایی سرکار خانم / جناب آقای دکتر حسن محمدزاده، مشاوره سرکار خانم / جناب آقای دکتر — و مشاوره سرکار خانم / جناب آقای دکتر — از آن دفاع شده است.

ماده ۳ به منظور جبران بخشی از هزینه های انتشارات دانشگاه، تعداد یک درصد شمارگان کتاب (در هر نوبت چاپ) را به «دفتر نشر آثار علمی» دانشگاه اهدا کند. دانشگاه می تواند مازاد نیاز خود را به نفع مرکز نشر در معرض فروش قرار دهد.

ماده ۴ در صورت عدم رعایت ماده ۳، ۵۰٪ بهای شمارگان چاپ شده را به عنوان خسارت به دانشگاه تربیت مدرس، تأدیه کند.

ماده ۵ دانشجوی تعهد و قبول می کند در صورت خودداری از پرداخت بهای خسارت، دانشگاه می تواند خسارت مذکور را از طریق مراجع قضایی مطالبه و وصول کند؛ به علاوه به دانشگاه حق می دهد به منظور استیفای حقوق خود، از طریق دادگاه، معادل وجه مذکور در ماده ۴ را از محل توقیف کتابهای عرضه شده نگارنده برای فروش، تأمین نماید.

ماده ۶ اینجانب علی محمد مصمم دانشجوی رشته ^{کارشناسی ارشد} _{مقطع کارشناسی ارشد} متعهد فوق و ضمانت اجرایی آن را قبول کرده، به آن ملتزم می شوم.

نام و نام خانوادگی: 

تاریخ و امضا: ۸۱،۲،۱۲

از انتشارات وزارت علوم ایران
تربیت مدرس

قدردانی

من لم یشکر المخلوق لم یشکر الخالق

حال که توفیق الهی نصیب گردید تا پایان نامه تحصیلی خود را ارائه دهم لازم است از کلبه‌ی افرادی که مرا در انجام این پروژه کمک نموده‌اند، خصوصاً استاد راهنمای گرامی جناب آقای دکتر محسن محمدزاده، استاد ممتحن و نماینده تحصیلات تکمیلی جناب آقای دکتر عین‌ا... پاشا و استاد ناظر آقای دکتر عباس گرامی تشکر کنم.

تقدیم به پدر و مادر بزرگوارم که سایه همتشان و وسعت
قناعتشان مامن آسایش من بوده است.
و تقدیم به روح پاک برادرم مهدی .

روشهای شناسایی داده‌های دورافتاده فضایی

چکیده

قبل از تحلیل آماری داده‌های فضایی برای شناخت بیشتر آنها تحلیل اکتشافی انجام می‌گیرد. داده‌های فضایی دارای نوعی وابستگی هستند که ناشی از موقعیت آنها در فضای مورد مطالعه می‌باشد. یک روش در تحلیل اکتشافی، شناسایی داده دورافتاده فضایی است. داده دورافتاده فضایی مشاهده‌ایست که نسبت به مقادیر همسایگی ناسازگار باشد، اما لزوماً با تمام مقادیر جامعه ناسازگار نیست. این داده‌ها اثرات نامطلوبی روی برآورد پارامترهای مدل‌های فضایی دارند. در این پایان‌نامه روشهای برآورد نیرومند فضایی و شناسایی داده‌های دورافتاده فضایی و مشکلاتی که در کاربرد آنها بوجود می‌آیند بررسی می‌شوند. وقتی بیش از یک داده دورافتاده وجود داشته باشد، بعلت اثرات درون‌آوری و برون‌بری این روشها در شناسایی داده دورافتاده موفق نیستند. لذا در این پایان‌نامه روشی برای شناسایی داده‌های دورافتاده فضایی ارائه می‌شود که با الگوریتم جستجوی پیشرو، مشاهدات را بر اساس سازگاریشان با مدل مشخصی مرتب می‌کند، و در نتیجه سبب شناسایی نقاط دورافتاده فضایی چندگانه می‌شود. به همین دلیل ابتدا روشهای پیش‌بینی فضایی ارائه شده است. در پایان مجموعه داده‌های میزان بارندگی در ایران بعنوان یک مثال کاربردی مورد تجزیه و تحلیل فضایی قرار می‌گیرد.

واژه‌های کلیدی : داده فضایی، تحلیل اکتشافی، کریگینگ، برآورد نیرومند، داده دورافتاده

فضایی، حذف موردی، جستجوی پیشرو، درون‌آوری و برون‌بری

فهرست مندرجات

۱	مقدمه	۱
۱	تاریخچه	۱.۱
۴	مفاهیم آمار فضایی	۲.۱
۷	تجزیه میدان تصادفی	۱.۲.۱
۷	تحلیل اکتشافی داده‌های فضایی	۳.۱
۱۰	داده‌های دورافتاده	۱.۳.۱
۱۲	منابع نقاط دورافتاده	۲.۳.۱
۱۳	اثرات درون آوری و برون بری	۳.۳.۱
۱۴	روشهای پیش‌بینی فضایی	۲
۱۴	مقدمه	۱.۲

۱۵	واریوگرام	۲.۲
۱۷	برآورد واریوگرام	۱.۲.۲
۱۷	روش گشتاوری	۲.۲.۲
۱۸	کریگینگ	۳.۲
۱۹	کریگینگ معمولی	۱.۳.۲
۲۲	کریگینگ عمومی	۲.۳.۲
۲۵	کریگینگ IRF_k	۳.۳.۲
۲۹	مدلهای خطی فضایی	۴.۲
۲۹	مدلهای فضایی با خطای اندازه‌گیری	۱.۴.۲
۳۱	ارزیابی پیش‌بینی فضایی با روش حذف موردی	۵.۲
۳۱	اثر حذف موردی بر ضرایب رگرسیونی و پیش‌بینی‌ها	۱.۵.۲

۳ روشهای پیش‌بینی نیرومند فضایی

۳۶	مقدمه	۱.۳
۳۶	مروری بر آماره‌های نیرومند	۲.۳
۳۷	میانگین پیراسته	۱.۲.۳
۳۸	برآوردگر نیرومند M	۲.۲.۳

۳۹	برآورد واریوگرام نیرومند	۳.۳
۴۰	تبدیلات اولیه	۱.۳.۳
۴۱	برآوردگر نیرومند میانه	۲.۳.۳
۴۵	کریگینگ نیرومند	۴.۳
۴۵	نقاط دورافتاده جدا شده	۱.۴.۳
۴۶	ویرایش داده‌ها	۲.۴.۳
۴۷	روش برآورد کریگینگ نیرومند	۳.۴.۳
۴۸	میانه وزنی	۴.۴.۳

۴ کشف نقاط دورافتاده فضایی

۴۹	مقدمه	۱.۴
۵۰	اثرات داده‌های دورافتاده بر روی واریوگرام	۲.۴
۵۲	داده‌های معدن طلا	۱.۲.۴
۵۴	شناسایی داده‌های دورافتاده	۳.۴
۵۵	نمودار پراکنش H	۱.۳.۴
۵۶	ماتریس عملگر H	۲.۳.۴
۵۷	فاصله ماهالانویس	۳.۳.۴
۵۸	روش توزیعی	۴.۳.۴

۵۸ خلاصه سازی سطری و ستونی	۴.۴
۵۹ پیراستن میانه	۱.۴.۴
۶۱ روش کرسی برای کشف داده دورافتاده	۲.۴.۴
۶۳ روش هاینینگ برای کشف داده دورافتاده	۳.۴.۴
۶۳ استفاده از ماتریس پیوند	۴.۴.۴
۶۳ روش هاوکینز	۵.۴.۴
۶۴ روشهای زمین آمار	۵.۴
۶۵ روش جک نایف	۱.۵.۴
۶۶ روش کریگینگ معمولی	۲.۵.۴
۶۷ روش کریگینگ IRF_k	۳.۵.۴
۶۷ روش باقیمانده‌ها	۶.۴
۶۸ اثرات درون آوری و برون بری	۷.۴
۷۱ روش هامپل	۱.۷.۴
۷۵ اندازه‌گیری اثرات درون آوری و برون بری	۲.۷.۴

۵ الگوریتم شناسایی داده دورافتاده فضایی

۷۶ مقدمه	۱.۵
----	-------------	-----

فهرست مندرجات

۵

۷۶	الگوریتم شناسایی داده‌های دورافتاده	۲.۵
۷۷	الگوریتم اصلی	۱.۲.۵
۷۸	انتخاب زیرمجموعه مقدماتی بر اساس روش اول	۲.۲.۵
۷۹	انتخاب زیرمجموعه مقدماتی بر اساس روش دوم	۳.۲.۵
۸۰	آزمون دورافتاده نبودن مشاهدات	۴.۲.۵
۸۰	تحلیلهای تشخیص نیرومند	۳.۵
۸۱	روش جستجوی پیشرو	۱.۳.۵
۸۳	جستجوی پیشرو با داده‌های خود همبسته فضایی	۴.۵
۸۶	تغییرات الگوریتم	۱.۴.۵
۸۸	الگوریتم جستجوی پیشرو بلوکی	۵.۵
۸۸	شناسایی نقاط دورافتاده فضایی	۱.۵.۵
۹۱	مثالهای عددی	۲.۵.۵
۹۴	شبیه سازی مونت کارلو	۳.۵.۵
۹۷	نتیجه‌گیری	۶.۵

۱۰۰

الف برنامه‌ها و خروجی‌های کامپیوتری

۱۰۰	الف.۱ محاسبه میانه پیراسته
-----	-------	----------------------------

فهرست مندرجات

و

الف.۲ شناسایی داده‌های دورافتاده فضایی ۱۰۱

الف.۳ تحلیل داده‌های بارندگی ۱۰۳

الف.۱.۳ محاسبه انواع واریوگرام ۱۰۴

الف.۲.۳ مقایسه واریوگرام تجربی و نیرومند ۱۰۴

۱۰۷

ب اثبات لم‌ها

ب.۱ لم ۱ ۱۰۷

ب.۲ لم ۲ ۱۰۸

ب.۳ لم (۳) ۱۰۸

ب.۱.۳ الف: ۱۰۸

ب.۲.۳ ب: ۱۰۸

ب.۴ اثبات لم (۴) ۱۰۹

ب.۵ لم (۵) ۱۰۹

فصل ۱

مقدمه

۱.۱ تاریخچه

در اغلب روشهای آمار کلاسیک فرض بر این است که مشاهدات مستقل از یکدیگرند، ولی در عمل با مواردی مواجه می‌شویم که مشاهدات به نوعی با یکدیگر وابسته‌اند. بعنوان مثال در داده‌های سریهای زمانی مشاهدات دارای همبستگی در طول زمان هستند. گاهی وابستگی مشاهدات، ناشی از موقعیت آنها در فضای مورد مطالعه می‌باشد. این گونه مشاهدات داده‌های فضایی^۱ نامیده می‌شود. نوعی از داده‌های فضایی که در موقعیتهای ثابت و مشخصی مشاهده می‌شوند را مشاهدات زمین آماری^۲ گویند. زمین آمار به شاخه‌ای از علم آمار گفته می‌شود که مبتنی بر نظریه متغیرهای ناحیه‌ای می‌باشد و نوعی دیگر از داده‌های فضایی که موقعیت مشاهده‌ای نیز یک متغیر تصادفی می‌باشد فرایندهای نقطه‌ای نامیده می‌شود.

اولین اظهارات آماری داده‌های فضایی به هالی (۱۶۸۶) بر می‌گردد که برای بررسی وزش بادهای موسمی نقشه‌های زمین را کنار هم گذاشته است. در دهه‌های اخیر نیز افراد

^۱ Spatial Data

^۲ Geostatistics

زیادی به نوعی با داده‌های فضایی سروکار داشته‌اند. استیودنت (۱۹۰۷) در بررسی توزیع ذرات معلق در مایعات پیشنهاد کرد که داده‌ها در سطح واحد جداگانه بدست آید. فیشر (۱۹۳۵) بوضوح از وابستگی فضایی در آزمایشهای زمین‌های کشاورزی آگاه بود. اما پایه‌های آمار فضایی به دنبال پژوهشهای قبلی توسط ماترون فرانسوی در سال ۱۹۶۲ بنا نهاده شد. در دهه ششم گاندین مباحث فضایی را در زمینه هواشناسی ارائه داده است. دیوید (۱۹۷۷) نیز در مورد پیش‌بینی‌های فضایی ذخایر معدن کار کرده است. و در نهایت کرسی (۱۹۹۳) با انتشار کتاب کاملی در تکامل این رشته نوپا قدم بزرگی را برداشت.

داده‌های دورافتاده مشاهداتی هستند که با دیگر مشاهدات سازگاری چندانی نداشته باشند. گاهی این مشاهدات باعث مشکلاتی در تجزیه و تحلیل داده‌ها می‌شوند و در بعضی از حالات نیز علاقمند به شناسایی داده‌های دورافتاده هستیم. بعنوان مثال در کارخانه‌های داروسازی علاقمند به شناسایی نوعی دارو هستیم که دارای اثر بسیار مثبتی روی بیماران است. در رگرسیون کلاسیک که مشاهدات مستقل از هم فرض شده‌اند، روشهای زیادی برای شناسایی داده‌های دورافتاده ارائه شده است. بویژه وقتی که از روشهای حداکثر درست‌نمایی برای برآورد پارامترهای رگرسیونی استفاده می‌شود، لازم است داده‌های دورافتاده مورد توجه قرار گیرند، زیرا این روش بشدت تحت تاثیر داده‌های دورافتاده قرار دارد. به همین علت روشهای جدیدی برای برآورد پارامترهای رگرسیونی ارائه شده است. از جمله می‌توان به روشهای نیرومند اشاره کرد که تا حد ممکن اثرات داده‌های دورافتاده را در برآورد پارامترهای رگرسیونی حداقل می‌کند. جنتلمن و ویلک (۱۹۷۵) داده‌های دورافتاده را زیر مجموعه‌ای از مشاهدات، که وقتی حذف شوند بیشترین کاهش را در مجموع مجدورات خطا دارند، تعریف نمودند.

در مورد داده دورافتاده کوتاه گودا (۱۹۸۴) ماکزیمم سیلابهای سالیانه را بررسی کرده است. همچنین موسسه آشناسی و مدیریت آب^۲ (IHII) نیز مبادرت به کشف نقاط