

پیش بینی ریسک بیماری لپتوسپیروز در شمال ایران با استفاده از مدل های حضور-شبه عدم حضور

مجتبی اصغرزاده^۱، علی اصغر آل شیخ^{۲*}، ابوالقاسم صادقی نیارکی^۳، یاسر ابراهیمیان^۴، رضا شیرزاد^۱

^۱ دانشجوی کارشناسی ارشد سیستم های اطلاعات مکانی - دانشکده مهندسی نقشه برداری - دانشگاه صنعتی خواجه-

نصیرالدین طوسی

m.a.neshelli@email.kntu.ac.ir

rezashirzad1994@gmail.com

^۲ استاد دانشکده مهندسی نقشه برداری - دانشگاه صنعتی خواجه نصیرالدین طوسی

alesheikh@kntu.ac.ir

^۳ استادیار دانشکده مهندسی نقشه برداری - دانشگاه صنعتی خواجه نصیرالدین طوسی

a.sadeghi@kntu.ac.ir

^۴ استادیار دانشکده عمران - دانشگاه صنعتی نوشیروانی بابل

y.ebrahimian@nit.ac.ir

(تاریخ دریافت آبان ۱۳۹۸، تاریخ تصویب بهمن ۱۳۹۸)

چکیده

لپتوسپیروز یا تب شالیزار یکی از بیماری های مشترک انسان و حیوان و با پراکندگی بالا در جهان است و به عنوان یک مشکل مهم بهداشت عمومی در ایران شناخته شده است. ترکیب علم همه گیرشناسی و سیستم اطلاعات مکانی این قابلیت را فراهم می نماید که بتوان مناطق تحت خطر بروز بیماری را مشخص نمود و با انجام فعالیت های پیشگیرانه همچون اطلاع رسانی و آموزش همگانی، بتوان از توسعه بیماری جلوگیری کرده و در نهایت آن را ریشه کن کرد. هدف این مطالعه بررسی کارایی روش های مختلف تولید داده های عدم حضور در مدل سازی بیماری لپتوسپیروز است که در تحقیقات پیشین در نظر گرفته نشده است تا در نهایت بتوان مدل سازی دقیق تری از شیوع این بیماری در استان های شمالی کشور به دست آورد. در این تحقیق از پنج روش متفاوت نقاط شبه عدم حضور تولید و با چهار روش شبکه عصبی مصنوعی، مدل تعمیم یافته خطی، جنگل تصادفی و گرادیان تقویتی مدل ریسک بیماری در منطقه مطالعه تولید شده است. نتایج نشان داده است که روش اعمال محدودیت فیزیکی با بافر به شعاع ۱۰ کیلومتر با مناسب ترین روش برای تولید داده های شبه عدم حضور بوده است. در نهایت مدل ایجاد شده که دارای بهترین ارزیابی در آماره TSS با مقادیر ۰،۷۶، ۰،۸۷، ۰،۸۴، ۰،۸۲، برای مدل های شبکه عصبی مصنوعی، مدل خطی تعمیم یافته، جنگل تصادفی، گرادیان تقویتی بوده به عنوان بهترین خروجی در نظر گرفته شده است.

واژگان کلیدی: لپتوسپیروز، تب شالیزار، سیستم اطلاعات مکانی، مدل سازی پراکنش بیماری، داده های شبه عدم حضور

* نویسنده رابط

۱- مقدمه

لپتوسپیروز^۱ (که در ایران به نام تب شالیزار شناخته می‌شود) از شایع‌ترین بیماری‌های مشترک بین انسان و حیوان، با پراکندگی بسیار بالا در جهان است [۱]. این بیماری همانند دیگر بیماری‌های زئونوز، با از بین بردن حیوانات، داشتن عوارض جدی بهداشتی در صورت ابتلا و احتمال فوت در انسان باعث خسارت‌های اقتصادی می‌شود. میزان مرگ‌ومیر گزارش‌شده از سراسر جهان برای این بیماری کمتر از ۵ تا ۳۰ درصد بوده است. در سال ۱۹۹۹ در کشور چین بیش از پانصد هزار مورد بیماری با مرگ‌ومیر ۷/۹ تا ۷۰/۹ درصد گزارش‌شده است، اما این ارقام چندان قابل‌اعتماد نیستند زیرا در بسیاری از مناطق این بیماری به‌خوبی به اثبات نمی‌رسد [۲]. لپتوسپیروز در مناطق گرمسیری، نیمه گرمسیری و معتدله، به‌ویژه در نواحی گرم و مرطوب شیوع بیشتری داشته و در اغلب این مناطق آندمیک است [۳]. این بیماری در کشورهای در حال توسعه مثل ایران یک بیماری شغلی محسوب می‌شود و بیشتر در کشاورزان، دامداران، کارکنان کشتارگاه‌ها، قصاب‌ها و ماهیگیران و در فصول گرم سال رخ می‌دهد [۲]. همچنین تشخیص تمایز آن با بسیاری از بیماری‌های دیگر در انسان مشکل است و نشانه‌های معمول آن در کل شبیه به آنفولانزا بوده و از نظر علائم بالینی نیز دارای طیف گسترده‌ای از موارد بدون علامت تا علائم کشنده و برق‌آسا است. ابتلا به این بیماری در تمام دوران بارداری برای مادر و جنین خطرناک است و می‌تواند یکی از علت‌های مهم مرگ جنین و نوزاد در مناطق آندمیک باشد [۴].

بیماری لپتوسپیروز یک بیماری نوظهور بوده که به دلیل وجود شرایط محیطی و اجتماعی مناسب برای شیوع در ایران، در سال‌های اخیر موارد زیادی از ابتلا به آن گزارش‌شده است [۵]. پیچیدگی تشخیص و خطر بالای ابتلا به آن در فصول گرم سال که معمولاً فصل کشاورزی و فعالیت اقتصادی برای ساکنین مناطق شمالی کشور بوده، باعث مشکلات مختلف اجتماعی و اقتصادی شده است.

وضعیت دانش و سطح اطلاع پزشکان عمومی که معمولاً اولین محل مراجعه بیماران مبتلا به لپتوسپیروز

است با توجه به شایع بودن این بیماری در مناطق شمالی کشور در سطح مناسبی قرار ندارد. در بررسی صورت گرفته توسط داودی و همکاران از ۴۰۰ پزشک عمومی در استان مازندران، سطح آگاهی ۶۲ درصد پزشکان شرکت‌کننده در مورد راه انتقال و پیش‌گیری این بیماری متوسط ارزیابی‌شده و سطح آگاهی ۹۲ درصد پزشکان از علائم آزمایشگاهی نیز ضعیف قلمداد شده است [۶].

به همین منظور بررسی مناطق تحت ریسک بیماری و پیش‌بینی مناطقی که امکان حضور بیماری در آن وجود دارد برای کنترل شیوع بیماری و افزایش سطح دانش مردم و پزشکان آن مناطق در مورد نحوه پیش‌گیری از ابتلا به این بیماری و علائم لپتوسپیروز و نحوه درمان ضروری به نظر می‌رسد.

هدف این تحقیق پیش‌بینی مناطق تحت ریسک بیماری لپتوسپیروز در سه استان شمالی ایران است که در این مناطق به‌عنوان یک بیماری آندمیک شناخته می‌شود. از آنجاکه اکثر روش‌های مدل‌سازی بیماری نیازمند ساختن داده‌های عدم حضور برای ساخت مدل هستند بررسی عملکرد روش‌های مختلف ایجاد نقاط شبه عدم حضور در مدل‌سازی این بیماری از اهداف دیگر این مطالعه است.

۲- پیشینه تحقیق

برخلاف دیدگاه عموم که همه‌گیرشناسی و بررسی ارتباط محیط به‌عنوان یک فاکتور در رخداد بیماری را یک‌رشته و خط تحقیقاتی جدید می‌دانند، سابقه این‌گونه فعالیت‌ها بسیار به عقب‌تر و به دوران کلاسیک یونانیان و روم برمی‌گردد. برای نمونه بقراط در نوشتار On Airs, Waters, and Places در حدود ۴۰۰ سال قبل از میلاد مسیح ایده آن‌که بیماری‌ها ممکن است با محیط فیزیکی مرتبط باشند را پروراند [۷]. تا قبل از قرن ۲۱، به‌طورمعمول همه‌گیرشناسی مکانی بیشتر به‌منظور تشخیص کلاسترهای مکانی به روش‌هایی مانند تحلیل نزدیک‌ترین همسایه، آماره شمارشی اتصالات، شاخص Moran، آزمون Knox و آزمون Cruzick به‌منظور ارزیابی توزیع غیر تصادفی افراد در معرض خطر یا بیماری استفاده می‌شد. از سال ۲۰۰۰ میلادی به بعد، روش‌هایی به کار گرفته شد که علاوه بر در نظر گرفتن توزیع غیر تصادفی افراد در معرض خطر، اطلاعات اضافی همبسته را نیز در

^۱ Leptospirosis

۲۰۱۶ به بررسی زمانی رابطه‌ی آن با میزان باران، رطوبت نسبی، شاخص پوشش گیاهی نرمال شده (NDVI)^۳، شاخص اصلاح شده آب (MNDWI)^۴ و دمای سطح زمین با استفاده از مدل خطی رگرسیون تعمیم داده شده پرداختند. طبق بررسی ایشان بروز این بیماری ارتباط قوی و معناداری با میزان باران و دما در طول سال دارد [۱۲].

در ایران نیز در چند سال اخیر تحقیقات مختلفی در زمینه بررسی پراکنش و مدل سازی بیماری ها و یا عوامل بیماری ها صورت گرفته است؛ برای نمونه، محمدی نیا و همکاران در سال ۲۰۱۷ با استفاده از روش رگرسیون وزن دار جغرافیایی (GWR)^۵ و بهره گیری از متغیرهای هواشناسی و توپوگرافی به مدل سازی همین بیماری در بین سال های ۲۰۰۹ تا ۲۰۱۱ در استان گیلان پرداختند. طبق نتایج ایشان، دما و رطوبت دارای بیشترین نقش در مدل سازی بوده اند [۱۳]. ملالو و همکاران در سال ۲۰۱۸ به مدل سازی اکولوژیکی یک نوع پشه خاکی به نام *Phlebotomus papatasi* که عامل انتقال لشمانيوز هست، در استان گلستان ایران پرداختند. ایشان ۱۴۲ نقطه در منطقه نمونه برداری کرده و در ۴۴ نقطه از آن این حشره وجود داشته است. با داشتن داده حضور و عدم حضور محققین این پژوهش از سه روش رگرسیون لجیستیکی، ماشین بردار پشتیبان و جنگل تصادفی^۶ و استفاده از روش اعتبار متقابل با تقسیم کردن داده ها به پنج قسمت برای مدل سازی استفاده کردند [۱۴]. آهنگرکانی و همکاران در سال ۲۰۱۹ به پیش بینی ریسک بیماری لپتوسپیروز در سه استان گیلان، مازندران و گلستان پرداختند. ایشان با استفاده از آماره Moran's I به این نتیجه رسیده اند که میزان رخداد این بیماری هم به صورت سالیانه و هم به صورت ماهیانه به صورت خوشه ای است و همچنین بیشترین خوشه های این بیماری در شهر یور و اردیبهشت وجود دارد که می تواند به دلیل تغییرات دمایی و یا فعالیت ویژه ای ساکنان منطقه در این دو ماه باشد. آن ها از دو مدل ماشین بردار پشتیبان و شبکه عصبی و از مقادیر بروز بین سال های ۲۰۰۹ تا ۲۰۱۳ برای کالیبره کردن مدل و از داده سال ۲۰۱۴ برای ارزیابی استفاده کردند. طبق

نظر می گرفت [۸]. برای نمونه، زیمس و همکاران در سال ۲۰۱۲ به مدل سازی هانتاویروس در سوئد پرداختند. در این پژوهش، از رخدادهای بین سال ۱۹۹۱ تا ۱۹۹۸ میلادی به عنوان موارد مورد مطالعه استفاده شده است. پژوهشگران این تحقیق از ۴ روش برای مدل سازی و مقایسه نتایج استفاده کرده اند. دو روش رگرسیون لجیستیکی^۱ و گرادیان تقویتی^۲ به عنوان مدل های حضور-عدم حضور، رگرسیون لجیستیکی اعتبار متقابل به عنوان مدل حضور-نقاط در دسترس و روش بیشینه آنتروپی یا MaxEnt به عنوان روش فقط حضور پیاده سازی شدند. طبق نظر نویسندگان این پژوهش، نتایج روش های متکی بر داده های فقط حضور مانند MaxEnt معمولاً دارای پیش بینی بیش از حد است [۹].

مسینا و همکاران در سال ۲۰۱۵ به بررسی مناطق تحت ریسک بیماری تب کریمه کنگو در جهان پرداخته اند. ایشان با استفاده از متغیرهای میانگین بارش، دمای سطح زمین، میانگین شاخص پوشش گیاهی و همچنین انحراف معیار آن (به عنوان شاخصی برای نمایان کردن تغییرات پوشش در سطح زمین)، نقشه پوشش گیاهی و با استفاده از روش درخت های تقویت شده رگرسیونی به مدل سازی این بیماری در سطح کره زمین پرداختند؛ طبق بررسی ایشان، ایران جزو پنج کشور با بیشترین مساحت تحت ریسک برای این بیماری است [۱۰].

ژائو و همکاران در سال ۲۰۱۶ با استفاده از دو روش MaxEnt و لجستیک رگرسیونی به پیش بینی ریسک بیماری لپتوسپیروز در کشور چین پرداختند. آن ها بیش از ۲۰۰۰ رخداد این بیماری را به واسطه آدرس آن ها ژئوکد کرده و از منطقه مطالعه به همان تعداد نقاط رخداد، نقاط شبه عدم حضور به صورت تصادفی انتخاب کردند. داده های مورد استفاده به عنوان ویژگی در مدل آن ها شامل متغیرهای محیطی و اجتماعی بوده است. طبق بررسی ایشان هر دو روش استفاده شده دارای قدرت پیش بینی بالا بوده و دو ویژگی محیطی میانگین دمای سالیانه و میزان بارش سالیانه در مدل سازی بیشترین نقش را ایفا کردند [۱۱]. دوانترا و همکاران در سال ۲۰۱۹ نیز با استفاده از داده های این بیماری در چین بین سال های ۲۰۰۶ تا

^۳ Normalized Difference Vegetation Index

^۴ Modified Normalized Difference Water Index

^۵ Geographically Weighted Regression

^۶ Random Forest

^۱ Logistic Regression

^۲ Boosted Regression Tree

نتیجه‌گیری ایشان میانگین دما و میزان بارش از متغیرهای دارای بیشترین نقش در مدل‌سازی بوده است [۵]. در این تحقیق سعی شده از روش‌های مختلف مدل‌سازی و با استفاده از داده‌های ۱۰ ساله رخداد بیماری برای تولید نقشه ریسک بیماری تب شالیزار در سه استان شمالی ایران بهره برده شود که در پژوهش‌های قبلی برای این بیماری استفاده نشده است. همچنین در این تحقیق اثرات روش‌های گوناگون پیاده‌سازی نقاط شبه عدم حضور روی مدل‌سازی ریسک بیماری مورد بررسی قرار گرفته شد که در پژوهش‌های پیشین در نظر گرفته نشده است.

۳- مواد و روش‌ها

در این بخش به تشریح منطقه مورد مطالعه، نحوه آماده‌سازی داده‌های بیماری، انتخاب متغیرها برای مدل‌سازی، نحوه ایجاد داده‌های شبه عدم حضور و در نهایت مدل‌های استفاده‌شده شرح داده شده است.

۳-۱- منطقه مورد مطالعه

مناطق ساحلی دریای خزر که در تقسیم‌بندی اقلیمی به دو قسمت شرقی و غربی (یا نواحی کرانه‌ای و پس‌کرانه‌ای خزری) و در تقسیم‌بندی سیاسی از سه استان گلستان، مازندران و گیلان تشکیل شده و به علت آنکه بیماری تب شالیزار در این مناطق شایع بوده، به‌عنوان

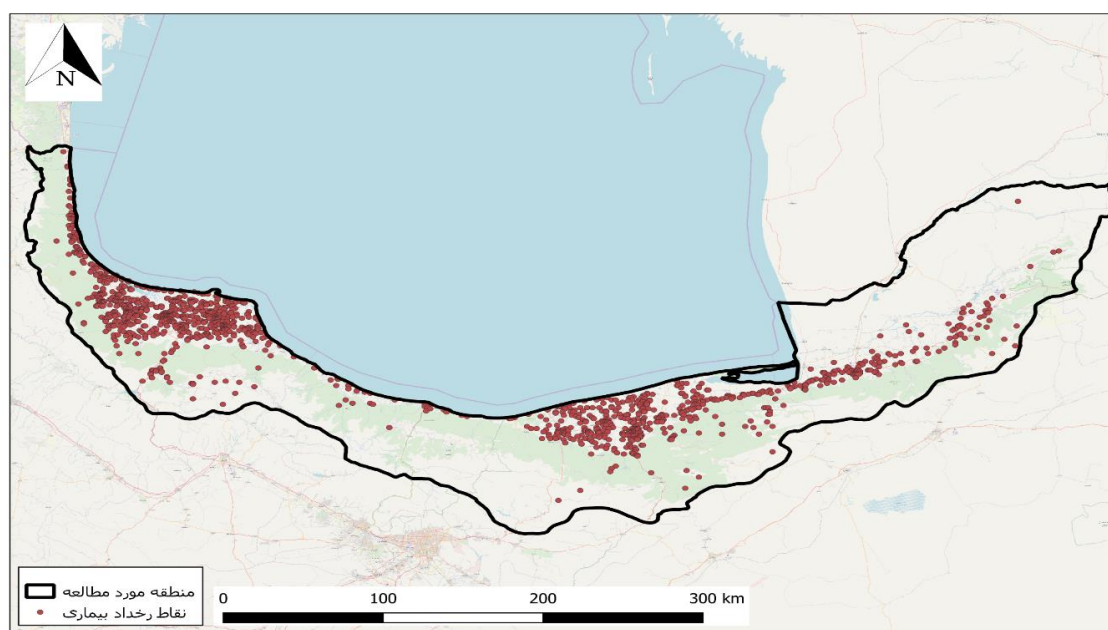
منطقه مورد مطالعه برگزیده شده است. منطقه مورد مطالعه در زون‌های ۳۹ و ۴۰ شمالی سیستم تصویر UTM و بین طول جغرافیایی ۴۸°۳۴ و ۵۶°۱۹ و همچنین عرض جغرافیایی ۴۶°۳۵ و ۳۸°۲۷ قرار گرفته است.

۳-۲- آماده‌سازی داده‌های بیماری

داده‌های مربوط به بیماران مبتلا به تب شالیزار در سه استان گلستان، مازندران و گیلان، از فروردین سال ۱۳۸۸ تا مرداد سال ۱۳۹۷ از مرکز کنترل و پیشگیری بیماری‌های واگیر وزارت بهداشت، درمان و آموزش پزشکی دریافت شده‌اند. مواردی که دارای سابقه‌ی مسافرت بوده‌اند و همچنین مواردی که دارای آدرس نبوده‌اند و یا آدرسشان قابل تشخیص و مکان‌یابی نبود حذف گشته‌اند؛ باقی موارد رخداد به‌واسطه آدرسشان، توسط Google Earth ژئوکد شدند. نقشه منطقه مورد مطالعه و پراکندگی نقاط در شکل ۱ قابل مشاهده است.

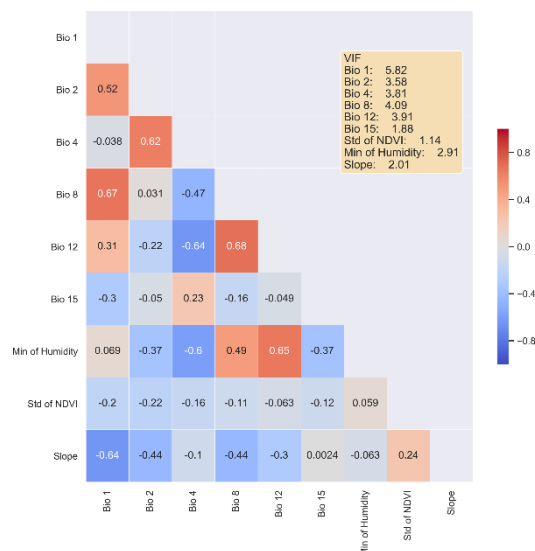
۳-۳- انتخاب متغیرهای مدل‌سازی

در مجموع ۲۶ متغیر محیطی در این تحقیق به‌عنوان متغیر اولیه مورد بررسی قرار گرفتند که شامل موارد زیر بودند:



شکل ۱- منطقه مورد مطالعه و پراکندگی نقاط رخداد بیماری لپتوسپیروز

جدول میزان همبستگی و رابطه هم‌خطی چندگانه متغیرهای انتخاب شده در شکل ۲ قابل مشاهده است.



شکل ۲- میزان همبستگی و رابطه هم‌خطی چندگانه متغیرهای انتخاب شده

از مزیت این روش انتخاب متغیر این است که مستقل‌ترین داده‌ها با کمترین همبستگی برای مدل‌سازی انتخاب می‌شوند و همچنین از وجود رابطه‌ی هم‌خطی چندگانه بالا نیز جلوگیری می‌شود. علاوه بر آن، انتخاب متغیرهای مدل‌سازی معنادار بوده و در این روش متغیرهای تأثیرگذار که با توجه به پژوهش‌های پیشین عملکردشان روی پراکنش بیماری به اثبات رسیده نیز حتماً در نظر گرفته می‌شوند.

۳-۴- ساختن داده‌های شبه عدم حضور

اکثر مدل‌هایی که برای بررسی مکان‌های مناسب پراکنش بیماری استفاده می‌شوند و نتایج قابل قبولی ارائه می‌کنند نیازمند به داده حضور - بیانگر نقاط رخداد بیماری- و داده عدم حضور (یا داده پس‌زمینه) - بیانگر نقاطی که بیماری وجود ندارد- هستند. داده‌های عدم حضور معمولاً توسط مشاهدات سامان‌مند و کلی در منطقه مطالعه به دست می‌آیند که می‌تواند هزینه‌بر باشد و قابلیت اجرا در مناطق وسیع را ندارد و در اکثر مواقع در دسترس نیست، در صورتی‌که داده‌های حضور از طریق مشاهدات مقطعی و گزارش‌ها رسیده شده از نهادها دریافت می‌شود [۱۷]، برای حل این مشکل روش‌های مختلفی برای ایجاد داده‌های شبه عدم‌حضور پیشنهاد شده است که با توجه به موضوع مطالعه و نوع داده‌ها متفاوت است [۱۸].

• ۱۹ متغیر هواشناسی مشتق شده از دما و بارش ماهیانه میانگین گرفته‌شده دیتاست worldclim 2.0 [۱۵]

• بیشینه، کمینه و میانگین رطوبت ماهیانه از ۱۳۸۶ تا ۱۳۹۷،

• نقشه ارتفاع و شیب استخراج‌شده توسط تصاویر ماهواره‌ای SRTM،

• نقشه کاربری زمین دریافت شده از سازمان جنگل‌ها، مراتع و آبخیزداری در شش کلاس کلی زراعت دیم، زراعت آبی و باغات، جنگل‌ها، مراتع، مناطق مسکونی و غیره.

سپس ابتدا دو متغیر میانگین دمای سالیانه و میزان بارش سالیانه که جزء مهم‌ترین متغیرها در تحقیقات پیشین شناخته‌شده بودند [۱۱، ۱۲، ۱۶] انتخاب‌شده و در ادامه متغیرهای پیوسته دیگر مجموعه متغیرهای اولیه در صورت داشتن شرایط زیر به انضمام لایه کاربری زمین، به لیست متغیرهای مدل‌سازی اضافه شدند:

۱. دارای کمترین همبستگی پیرسان^۱ با متغیرهای دیگر باشند

۲. با اضافه شدن این متغیر کمترین میزان رابطه هم‌خطی چندگانه ایجاد شود

روند اضافه شدن متغیرها تا زمانی که میزان همبستگی پیرسان از حد آستانه ۰٫۷ و میزان رابطه هم‌خطی چندگانه از حد آستانه ۸ عبور نکند ادامه یافت. در پایان متغیرهای انتخاب‌شده شامل موارد زیر شدند:

جدول ۱- متغیرهای انتخاب‌شده برای مدل‌سازی

نام متغیر	توضیح
Bio1	میانگین دمای سالیانه
Bio2	متوسط دامنه روزانه دما
Bio4	تغییرات دمای فصلی
Bio8	میانگین دما در پربارش‌ترین فصل سال
Bio12	میزان بارش سالیانه
Bio15	تغییرات بارش فصلی
Min of humidity	کمینه رطوبت ماهیانه
Std of NDVI	انحراف معیار شاخص پوشش گیاهی
Slope	شیب

^۱ Pearson correlation

در این تحقیق از پنج روش برای تولید نقاط شبه عدم حضور استفاده شده است و سپس به مقایسه بررسی نتایج حاصل پرداخته شده است:

۱. نقاط به صورت تصادفی در منطقه انتخاب شده‌اند (RND).
 ۲. یک بافر به شعاع ۵ کیلومتر دور نقاط حضور پیاده شده و نقاط شبه عدم حضور در خارج این مناطق به صورت تصادفی انتخاب شده‌اند (BUFFER-5).
 ۳. همانند روش دوم ولی با این تفاوت که بافر به شعاع ۱۰ کیلومتر انتخاب شده است (BUFFER-10).
 ۴. با توجه به روش پیشنهاد شده در [۱۹] ابتدا یک مدل طبقه‌بندی یک-کلاسه ماشین‌بردار پشتیبان پیاده شده است. سپس از مناطقی که این مدل نامناسب تشخیص داده نقاط شبه عدم حضور به صورت تصادفی انتخاب شده است (SVM).
 ۵. ابتدا یک مدل کلاسیک توزیع گونه‌ها به نام بایوکلیم پیاده شده و سپس از مناطقی که این مدل نامناسب تشخیص داده شده است نقاط شبه عدم حضور انتخاب شده است (BIOCLIM).
- در این پژوهش برای ساده‌تر کردن پیاده‌سازی و آنالیز نتایج، تعداد نقاط عدم-حضور برابر با تعداد نقاط حضور انتخاب شده است و همچنین برای هر روش ۵۰ دسته داده تولید شده و مدل‌سازی شده است.

۳-۵- مدل‌های استفاده شده

در این پژوهش برای مدل‌سازی بیماری لپتوسپیروز در منطقه مورد مطالعه از چهار روش شبکه عصبی مصنوعی، مدل خطی تعمیم‌یافته^۱، جنگل تصادفی و گرادیان تقویتی استفاده شده است که در ادامه به توضیح آنان پرداخته می‌شود.

۳-۵-۱- شبکه عصبی مصنوعی

شبکه‌های عصبی مصنوعی، سامانه‌ها و روش‌های محاسباتی نوینی برای یادگیری ماشینی، نمایش دانش و در انتها اعمال دانش به دست‌آمده در جهت پیش‌بینی پاسخ‌های خروجی از سامانه‌های پیچیده هستند. ایده

اصلی این گونه شبکه‌ها الهام گرفته از شیوه کارکرد سیستم عصبی زیستی و سامانه عصبی مغز انسان برای پردازش داده‌ها و اطلاعات برای یادگیری و ایجاد دانش است. عنصر کلیدی این ایده، ایجاد ساختارهایی جدید برای سامانه پردازش اطلاعات است. این سیستم از شمار زیادی عناصر پردازشی فوق‌العاده به هم پیوسته با نام نورون تشکیل شده که برای حل یک مسئله باهم هماهنگ عمل می‌کنند و سپس توسط سیناپس‌ها^۲ (ارتباطات الکترومغناطیسی) اطلاعات را منتقل می‌کنند. یک شبکه عصبی را می‌توان به عنوان یک گراف وزن‌دار جهت‌دار در نظر گرفت که نورون‌ها به عنوان گره‌ها و ارتباط بین نورون‌ها و یال‌های جهت‌دار مشخص می‌گردد [۲۰].

۳-۵-۲- مدل خطی تعمیم‌یافته

مدل‌های خطی تعمیم‌یافته همان روش‌های رگرسیون ساده کمترین مربعات هستند که برای مدل‌سازی روابط اکولوژیکی توسعه داده شده‌اند و قابلیت سازگار شدن با دامنه منعطفی از ساختارهای خطا هستند. مدل‌های خطی تعمیم‌یافته بر اساس فرض یک رابطه (تابع ارتباط^۳) بین مقدار خروجی و یک ترکیب خطی از متغیرهای توصیفی بنانهاده شده‌اند. این رابطه، یک تناظر میان رخدادها و هر متغیر محیطی تعریف می‌کند و توزیع خطا نیز می‌تواند از حالت‌های مختلف همچون نرمال، دوتایی، پوی سون و توزیع منفی دوتایی انتخاب شود [۲۱]. استفاده از این ارتباط باعث خطی شدن و محدود نگه‌داشتن پیش‌بینی‌ها در یک بازه مربوط به مقادیر توصیفی است. برخلاف رگرسیون‌های کمترین مربعات که می‌توانند مقادیر غیرممکن نیز پیش‌بینی کنند (همچون احتمال بیش از یک و یا مقادیر منفی)، پیش‌بینی‌های مدل‌های خطی تعمیم‌یافته در محدوده مقادیر مشاهداتی است (حضور-عدم حضور [0,1] و مقادیر احتمال بین این دو حد) [۲۲].

۳-۵-۳- جنگل تصادفی

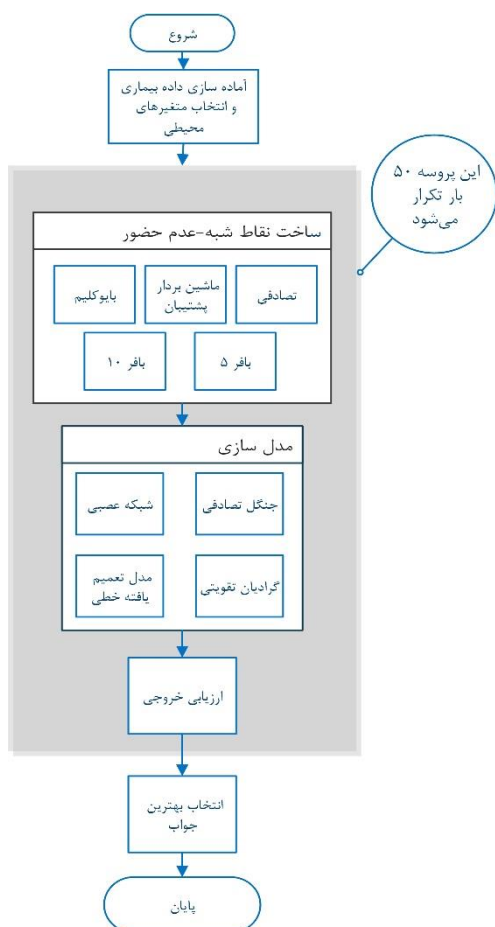
جنگل تصادفی یک از روش‌های یادگیری گروهی است. در زمان ساخت اجزای درخت تصمیم، در هر مرحله از تصمیم‌گیری برای جداسازی، الگوریتم اول یک زیرمجموعه

^۲ Synapse

^۳ Link function

^۱ Generalized linear model (GLM)

است و میزان خطا در آن معمولاً از طریق مجذور اختلاف مقدار مشاهده شده و پیش‌بینی شده محاسبه می‌شود [۲۷]. در شکل ۳ روند کلی پژوهش نشان داده شده است.



شکل ۳- روند مراحل مدل سازی

۳- پیاده سازی و نتایج

مدل ماشین بردار پشتیبان برای ایجاد نقاط شبه عدم حضور در زبان برنامه نویسی Python و با استفاده از کتابخانه Scikit-learn انجام شده است [۲۸]. مدل بایوکلیم نیز در زبان برنامه نویسی R و با استفاده از کتابخانه dismo انجام شده است [۲۹]. مدل سازی های توزیع بیماری در زبان برنامه نویسی R و با استفاده از کتابخانه BIOMOD2 صورت گرفته است [۳۰].

برای کاهش اثر خودهمبستگی داده ها در ارزیابی و همچنین به دلیل آن که روش اعتبارسنجی معمولی در محیط ساختاریافته باعث امتیاز بیش از حد در ارزیابی می شود از روش اعتبار سنجی مکانی استفاده شده است [۳۱]. این روش بدین صورت بوده که منطقه در راستای طول جغرافیایی به پنج قسمت تقسیم شده و به صورت

تصادفی از داده های اصلی مشتق می کند و سپس فرایند کلاسه بندی در زیرمجموعه ایجاد شده انجام می شود؛ در آخر جواب نهایی از مجموعه جواب ها به دست می آید [۲۳].

الگوریتم جنگل تصادفی به طور خلاصه به صورت زیر عمل می کند:

۱. n_{tree} زیرمجموعه ی خودراه انداز^۱ از مجموعه داده اصلی استخراج می شود.

۲. برای هر یک از نمونه ها، یک کلاسه بندی اصلاح نشده^۲ یا یک درخت رگرسیون با تغییر زیر انجام می شود: در هر نقطه، به جای انتخاب بهترین تقسیم بین ویژگی ها، به طور تصادفی از ویژگی ها k نمونه گرفته شده و بهترین حالت تقسیم از بین آن ها انتخاب می شود.

۳. پیش بینی داده جدید با جمع کردن پیش بینی های n_{tree} درخت (رای اکثریت برای کلاسه بندی، میانگین گیری برای رگرسیون) [۲۴، ۲۵].

۳-۵-۴- گرادینان تقویتی

درخت های تقویت شده رگرسیونی یک متد گروهی است که سعی در برازش دادن مدل های آماری است که در آن از ترکیب دو الگوریتم درخت های رگرسیون^۳ و روش های تقویتی^۴ بهره برده می شود و در پایان گرادینان تقویتی می تواند یک مدل رگرسیون افزایشی در نظر گرفته شود که هر ترم آن را می توان به عنوان یک درخت ساده در نظر گرفت که به صورت مرحله به مرحله اضافه شده است [۲۶].

آنالیزهای درخت رگرسیون و کلاسه بندی نوعی از متدولوژی درخت تصمیم گیری هستند که در آن ها فضای داده به صورت بازگشتی^۵ تقسیم شده و یک مدل ساده برای پیش بینی به آن برازش می شود و این تقسیم بندی در آخر می تواند به صورت یک درخت به نمایش دربیاید. درخت های کلاسه بندی برای داده های وابسته به هم که تعداد محدودی از اعداد بدون نظم را دربر می گیرند مناسب بوده که در آن خطای پیش بینی به صورت تعداد کلاسه بندی اشتباه اندازه گیری می شود؛ درخت های رگرسیون برای داده های وابسته به هم که مقادیرشان به صورت پیوسته یا مقادیر گسسته ولی منظم، مناسب

^۱ bootstrap

^۲ Unpruned classification

^۳ Regression trees

^۴ boosting

^۵ recursive

اعتبار سنجی متقابل هربار از چهار قسمت برای مدل‌سازی و یک قسمت برای ارزیابی استفاده شده است.

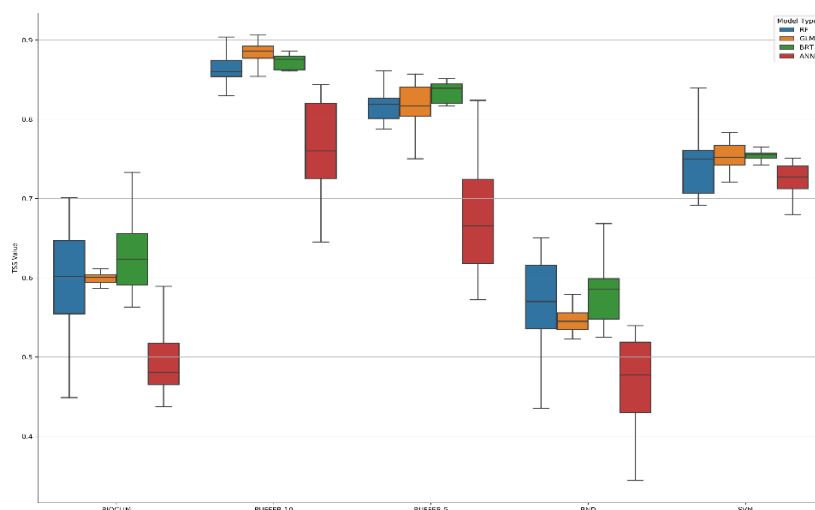
ارزیابی مدل‌ها توسط آماره TSS انجام گرفته است که نمایانگر میانگین نرخ پیش‌بینی موفقیت‌آمیز صحیح در ایستگاه‌های حاضر و غایب تعریف می‌شود. استفاده از این آماره به جای آماره Kappa برای بررسی قدرت مدل‌هایی که پیش‌بینی حضور-عدم حضور تولید می‌کنند پیشنهاد شده است [۳۲].

نمودارهای جعبه‌ای در شکل ۴ بیانگر این هستند که در بین پنج روش ایجاد نقاط شبه عدم حضور، روشی که در آن از بافر به شعاع ۱۰ کیلومتر استفاده شده دارای جواب‌های بهتری نسبت به سایر روش‌ها است.

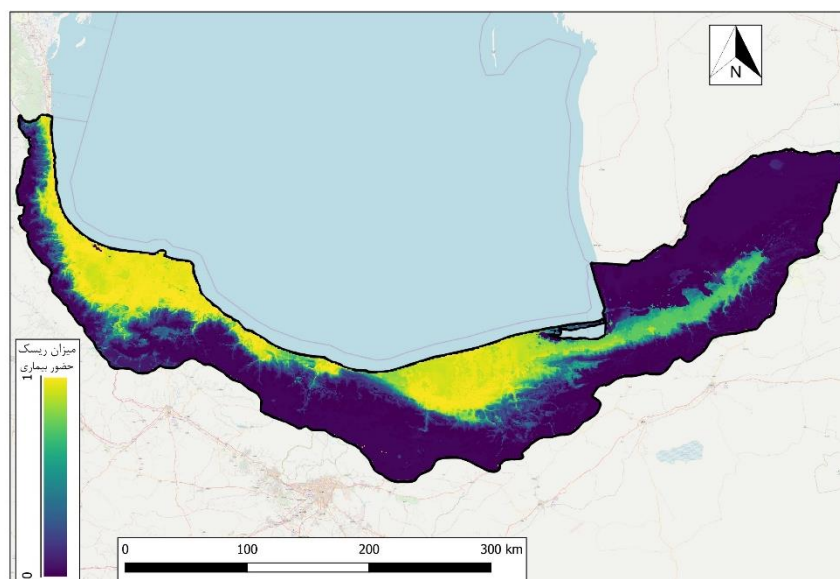
در ادامه از میان خروجی‌های روش بافر ۱۰ کیلومتر، مدل با بالاترین میانگین ارزیابی TSS به ازای الگوریتم‌های پیاده شده به عنوان خروجی نهایی برگزیده شد و نقشه نهایی ریسک با گرفتن میانگین وزنی از چهار مدل شبکه عصبی مصنوعی، مدل خطی تعمیم‌یافته، جنگل تصادفی و گرادیان تقویتی تهیه شد. مقدار آماره TSS و Kappa برای خروجی نهایی در جدول زیر آمده شده‌اند:

جدول ۲- نتایج ارزیابی مدل نهایی

نوع ارزیابی	شبکه عصبی مصنوعی	مدل خطی تعمیم‌یافته	جنگل تصادفی	گرادیان تقویتی
TSS	۰,۸۳	۰,۸۹	۰,۸۸	۰,۸۸
KAPPA	۰,۷۶	۰,۸۷	۰,۸۴	۰,۸۲



شکل ۴- نمودار نتایج مدل‌های اجرا شده برای هر مدل و با توجه به الگوریتم ایجاد نقاط شبه عدم حضور استفاده شده (انواع مدل به ترتیب RF: جنگل تصادفی، GLM: مدل خطی تعمیم‌یافته، BRT: گرادیان تقویتی، ANN: شبکه عصبی مصنوعی)



شکل ۵- نقشه ریسک لپتوسپیروز در منطقه مورد مطالعه

طبق نقشه نهایی (شکل ۵)، به‌طور کلی بخش‌های مرکزی استان گیلان و مازندران دارای بیشترین خطر بروز این بیماری هستند. بعد از میانگین‌گیری نقاط ریسک در سطح دهستان‌ها، برای استان گیلان، دهستان‌های لیل، اسالم، اسلام‌آباد، چهارده و لغمجان نسبتاً دارای بالاترین میزان در کل منطقه مورد مطالعه بودند.

برای استان مازندران دهستان‌های لیتکوه آمل، بابل‌کنار بابل، کوهساران و بالا تجن قائم‌شهر نسبتاً دارای بالاترین میزان ریسک در این استان بوده‌اند. مناطق تحت ریسک برای استان گلستان نیز شامل دهستان‌های بنفشه تپه، لیوان، استرآباد شمالی و انزان هستند.

اثرگذارترین متغیرهای محیطی محاسبه‌شده برای مدل شبکه عصبی مصنوعی به ترتیب برابر با میزان بارش سالیانه، کاربری زمین، کمینه رطوبت و دمای سالیانه؛ برای مدل تعمیم‌یافته خطی کمینه رطوبت ماهیانه، دمای سالیانه، کاربری زمین و تغییرات دمای فصلی، برای مدل جنگل تصادفی میزان بارش سالیانه، کاربری زمین، کمینه رطوبت ماهیانه و تغییرات دمای فصلی و برای مدل گرادیان تقویتی میزان بارش سالیانه، کاربری زمین، کمینه رطوبت ماهیانه، و تغییرات دمای فصلی بدست آمده‌اند. به طور کلی میزان بارش سالیانه را می‌توان اثرگذارترین متغیر در این مدل‌سازی در نظر گرفت که با توجه به بیماری لپتوسپیروز و نحوه شیوع آن که معمولاً توسط آب آلوده صورت می‌گیرد [۱] توجیه پذیر است.

۴- بحث و نتیجه‌گیری

به دلیل آن‌که کیفیت داده‌های حضور و عدم حضور برای رسیدن به یک مدل با دقت مناسب بسیار مهم است، یافتن روش مناسب برای الگوسازی داده‌های شبه عدم حضور حیاتی است. روش‌های کلی برای ایجاد این داده‌ها را به‌طور کلی می‌توان به سه دسته تقسیم‌بندی کرد: (۱) تولید تصادفی نقاط در داده‌های پس‌زمینه، (۲) تولید نقاط با استفاده از قرار دادن محدودیت فاصله جغرافیایی از نقاط معلوم و (۳) تولید نقاط شبه عدم حضور در مناطقی که از نظر محیطی متفاوت با مناطق حضور هستند [۱۹]. در این تحقیق از پنج روش برای تولید این نقاط استفاده شد که دو روش اعمال بافر ۵ و ۱۰ کیلومتر نماینده محدودیت‌های فاصله جغرافیایی و دو روش استفاده از ماشین‌بردار پشتیبان و مدل بایوکلیم نماینده محدودیت‌های محیطی بودند.

روش تولید نقاط شبه عدم حضور به‌واسطه پیاده‌سازی ماشین‌بردار پشتیبان باوجود داشتن کمترین انحراف معیار در نمودار ارزیابی شکل ۴ که بیانگر توانایی بالای این روش و روش‌های شبیه به آن (روش‌های تشخیص داده‌های خارج از محدوده نظارت‌نشده) در ارائه نتایج یکدست است اما نسبت به روش‌هایی که فقط محدودیت مکانی برای تولید نقاط شبه عدم حضور برقرار می‌کنند دارای توانایی کمتری است. از طرف دیگر، در ارزیابی MPA^۱ (حداقل سطح به‌دست‌آمده با در نظر گرفتن حد آستانه‌ای که در آن بیش از ۹۰ درصد پیش‌بینی‌های نقاط رخداد بالاتر از این حد باشند [۳۳]) مدل ساخته‌شده از این روش دارای بیشترین مقدار این آماره بوده که نشان‌دهنده برآورد بیش‌ازحد این مدل نسبت به سایر روش‌های پیاده شده است.

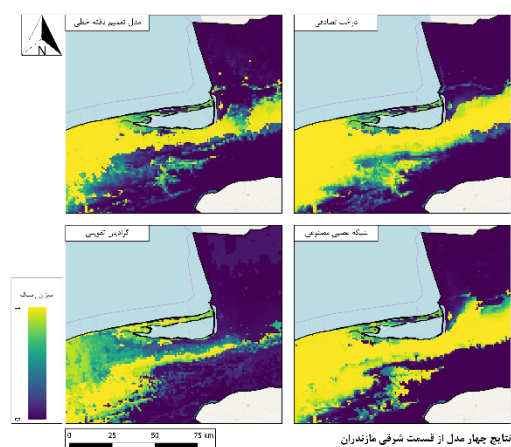
مدل بایوکلیم یک مدل کلاسیک توزیع گونه بوده که فقط از داده حضور برای مدل‌سازی استفاده می‌کند و یک ابر فضای n-بعدی را با استفاده از محدودیت‌های ساده تعریف می‌کند. این روش به دلیل قدرت کم در مدل‌سازی، حساب نکردن ارتباط بین پارامترهای محیطی و عدم توانایی در استفاده از داده‌های طبقه‌بندی دارای توانایی محدودی در مقایسه با روش‌های مدل‌سازی دیگر که از داده حضور-عدم حضور استفاده می‌کنند است [۳۴].

همچنین در این روش، به علت مدل‌سازی پایین‌تر از حد انتظار، فضای تخصیص‌یافته شده برای انتخاب نقاط شبه عدم حضور تقریباً بزرگ و حدوداً شامل کل منطقه مورد مطالعه بوده است. در نتیجه همان‌طور که در نمودارهای شکل ۶ مشخص است، این روش مزیت چشمگیری نسبت به روش تصادفی ندارد.

روش انتخاب نقاط شبه عدم حضور کاملاً تصادفی که از ساده‌ترین و پرکاربردترین روش‌ها برای تولید نقاط است [۳۵]، دارای ضعیف‌ترین عملکرد و بالاترین انحراف معیار در ارزیابی است، نقشه ریسک محاسبه‌شده به‌واسطه این روش دارای کمترین MPA است که بیان‌گر آن است این روش نسبت به سایر روش‌ها مناطق کمتری را به‌عنوان مناطق ریسک شناخته است.

از پنج روش پیاده شده، دو روشی که فقط محدودیت فاصله به‌واسطه بافر اعمال‌شده است، دارای عملکرد بهتری

^۱ Minimal Predicted Area



شکل ۶- پیش‌بینی چهار مدل برای بخشی از منطقه مورد مطالعه

۵- پیشنهادها

مطالعات آینده می‌تواند نقش اثرات همبستگی مکانی در مدل‌سازی‌های بیماری را در نظر بگیرد و اثر روش‌های مختلف همچون نمونه‌برداری از داده‌ها و روش‌های اتورگرسیون را روی مدل‌سازی مورد بررسی قرار دهد. همچنین پژوهش‌های آینده می‌تواند به بررسی عملکرد مدل‌هایی که فقط از داده حضور استفاده می‌کنند (مانند MaxEnt) و مقایسه آن با مدل‌های حضور-عدم حضور بپردازند.

نقش اثرگذار متغیرهای محیطی و کاربری زمین در توزیع این پژوهش نشان داده شده و این بیماری در منطقه شمال کشور مدل‌سازی شده است. با توجه به تغییرات آب و هوایی، روند گرم شدن زمین و همچنین تغییرات پوشش گیاهی و کاربری زمین، بررسی اثرات این متغیرها روی شیوع این بیماری و یا گسترش این بیماری در مناطق دیگر می‌تواند موضوع تحقیق و بررسی جداگانه قرار گیرد.

سپاسگزاری

در پایان، نگارندگان بر خود لازم می‌دانند از مدیریت و کارکنان بخش مرکز کنترل و پیشگیری بیماری‌های واگیر وزارت بهداشت، درمان و آموزش پزشکی ایران بابت کمک در تأمین داده‌های بیماری لپتوسپیروز تشکر نمایند.

مراجع

- [1] F. Costa et al., "Global morbidity and mortality of leptospirosis: a systematic review," PLoS neglected tropical diseases, vol. 9, no. 9, p. e0003898, 2015.
- [2] A. Rafiei, A. Hedayati Zadeh omran, F. Babamahmoodi, R. Alizqadeh Navaee, and R. Valadan, "Review of Leptospirosis in Iran," (in eng), Journal of Mazandaran University of Medical Sciences, Research(Original) vol. 22, no. 94, pp. 102-110, 2012.
- [3] A. R. Bharti et al., "Leptospirosis: a zoonotic disease of global importance," The Lancet infectious diseases, vol. 3, no. 12, pp. 757-771, 2003.

- [4] P. Rezasoltani and F. Rafat, "Leptospirosis and its consequences in pregnancy," (in Fa), Journal of Holistic Nursing and Midwifery, vol. 19, no. 1, pp. 37-, 2009.
- [5] M. Ahangarcani, M. Farnaghi, M. R. Shirzadi, P. Pilesjö, and A. Mansourian, "Predictive risk mapping of human leptospirosis using support vector machine classification and multilayer perceptron neural network," Geospatial health, vol. 14, no. 1, 2019.
- [6] A. Davoodi, F. Babamahmoodi, M. Khademloo, A. Alikhani, and N. Najafi, "The Awareness of General Practitioners of Mzandaran Province of Leptospirosis and its Related Factors," (in eng), Journal of Mazandaran University of Medical Sciences, Research(Original) vol. 22, no. 93, pp. 98-102, 2012.
- [7] R. H. Friis and T. A. Sellers, Epidemiology for public health practice, 5th ed. Burlington, Mass.: Jones & Bartlett Learning, 2014, pp. xxiv, 804 p.
- [8] T. Carpenter, "The spatial epidemiologic (r) evolution: a look back in time and forward to the future," Spatial and Spatio-temporal Epidemiology, vol. 2, no. 3, pp. 119-124, 2011.
- [9] C. B. Zeimes, G. E. Olsson, C. Ahlm, and S. O. Vanwambeke, "Modelling zoonotic diseases in humans: comparison of methods for hantavirus in Sweden," International journal of health geographics, vol. 11, no. 1, p. 39, 2012.
- [10] J. P. Messina et al., "The global distribution of Crimean-Congo hemorrhagic fever," Transactions of the Royal Society of Tropical Medicine and Hygiene, vol. 109, no. 8, pp. 503-513, 2015.
- [11] J. Zhao et al., "Mapping risk of leptospirosis in China using environmental and socioeconomic data," BMC infectious diseases, vol. 16, no. 1, p. 343, 2016.
- [12] P. W. Dhewantara et al., "Climate variability, satellite-derived physical environmental data and human leptospirosis: A retrospective ecological study in China," Environmental research, 2019.
- [13] A. Mohammadinia, A. Alimohammadi, and B. Saeidian, "Efficiency of geographically weighted regression in modeling human leptospirosis based on environmental factors in Gilan Province, Iran," Geosciences, vol. 7, no. 4, p. 136, 2017.
- [14] A. Mollalo, A. Sadeghian, G. D. Israel, P. Rashidi, A. Sofizadeh, and G. E. Glass, "Machine learning approaches in GIS-based ecological modeling of the sand fly *Phlebotomus papatasi*, a vector of zoonotic cutaneous leishmaniasis in Golestan province, Iran," Acta tropica, vol. 188, pp. 187-194, 2018.
- [15] S. E. Fick and R. J. Hijmans, "WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas," International journal of climatology, vol. 37, no. 12, pp. 4302-4315, 2017.
- [16] M. AhangarCani, M. Farnaghi, and M. Shirzadi, "Predictive Map of Spatio-Temporal Distribution of Leptospirosis Using Geographical Weighted Regression and Multilayer Perceptron Neural Network Methods," Journal of Geomatics Science and Technology, vol. 6, no. 2, pp. 79-98, 2016.
- [17] K. B. Stevens and D. U. Pfeiffer, "Spatial modelling of disease using data-and knowledge-driven approaches," Spatial and spatio-temporal epidemiology, vol. 2, no. 3, pp. 125-133, 2011.
- [18] J. M. Lobo, A. Jiménez-Valverde, and J. Hortal, "The uncertain nature of absences and their importance in species distribution modelling," Ecography, vol. 33, no. 1, pp. 103-114, 2010.
- [19] S. D. Senay, S. P. Worner, and T. Ikeda, "Novel three-step pseudo-absence selection technique for improved species distribution modelling," PLoS One, vol. 8, no. 8, p. e71218, 2013.
- [20] M. T. Hagan, H. B. Demuth, M. H. Beale, and O. De Jesús, Neural network design. Pws Pub. Boston, 1996.
- [21] A. T. Peterson et al., Ecological niches and geographic distributions (MPB-49). Princeton University Press, 2011.
- [22] A. Guisan and N. E. Zimmermann, "Predictive habitat distribution models in ecology," Ecological modelling, vol. 135, no. 2-3, pp. 147-186, 2000.
- [23] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5-32, 2001.
- [24] A. Liaw and M. Wiener, "Classification and regression by randomForest," R news, vol. 2, no. 3, pp. 18-22, 2002.
- [25] H. Trevor, T. Robert, and F. JH, "The elements of statistical learning: data mining, inference, and prediction," ed: New York, NY: Springer, 2009.
- [26] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," Journal of Animal Ecology, vol. 77, no. 4, pp. 802-813, 2008.
- [27] W. Y. Loh, "Classification and regression trees," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 1, no. 1, pp. 14-23, 2011.

- [28] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825-2830, 2011.
- [29] R. J. Hijmans, S. Phillips, J. Leathwick, J. Elith, and M. R. J. Hijmans, "Package 'dismo'," *Circles*, vol. 9, no. 1, pp. 1-68, 2017.
- [30] W. Thuiller, D. Georges, R. Engler, F. Breiner, M. D. Georges, and C. W. Thuiller, "Package 'biomod2'," *Species distribution modeling within an ensemble forecasting framework* <https://CRAN.R-project.org/package=biomod2>, 2016.
- [31] R. Valavi, H. Shafizadeh-Moghadam, A. Matkan, A. Shakiba, B. Mirbagheri, and S. H. Kia, "Modelling climate change effects on Zagros forests in Iran using individual and ensemble forecasting approaches," *Theoretical and Applied Climatology*, vol. 137, no. 1-2, pp. 1015-1025, 2019.
- [32] O. Allouche, A. Tsoar, and R. Kadmon, "Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS)," *Journal of applied ecology*, vol. 43, no. 6, pp. 1223-1232, 2006.
- [33] R. Engler, A. Guisan, and L. Rechsteiner, "An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data," *Journal of applied ecology*, vol. 41, no. 2, pp. 263-274, 2004.
- [34] T. H. Booth, H. A. Nix, J. R. Busby, and M. F. Hutchinson, "BIOCLIM: the first species distribution modelling package, its early applications and relevance to most current MAXENT studies," *Diversity and Distributions*, vol. 20, no. 1, pp. 1-9, 2014.
- [35] M. Iturbide, J. Bedia, S. Herrera, O. del Hierro, M. Pinto, and J. M. Gutiérrez, "A framework for species distribution modelling with improved pseudo-absence generation," *Ecological Modelling*, vol. 312, pp. 166-174, 2015.