عنوان سخنرانی:

مدلسازی داده های نامتوازن مکانی با استفاده از دو رویکرد سطح داده و سطح الگوریتم از طریق مدلهای یادگیری ماشین: مطالعه موردی مدلسازی رشد فیزیکی شهر

ارائه دهنده:

محمد احمدلو

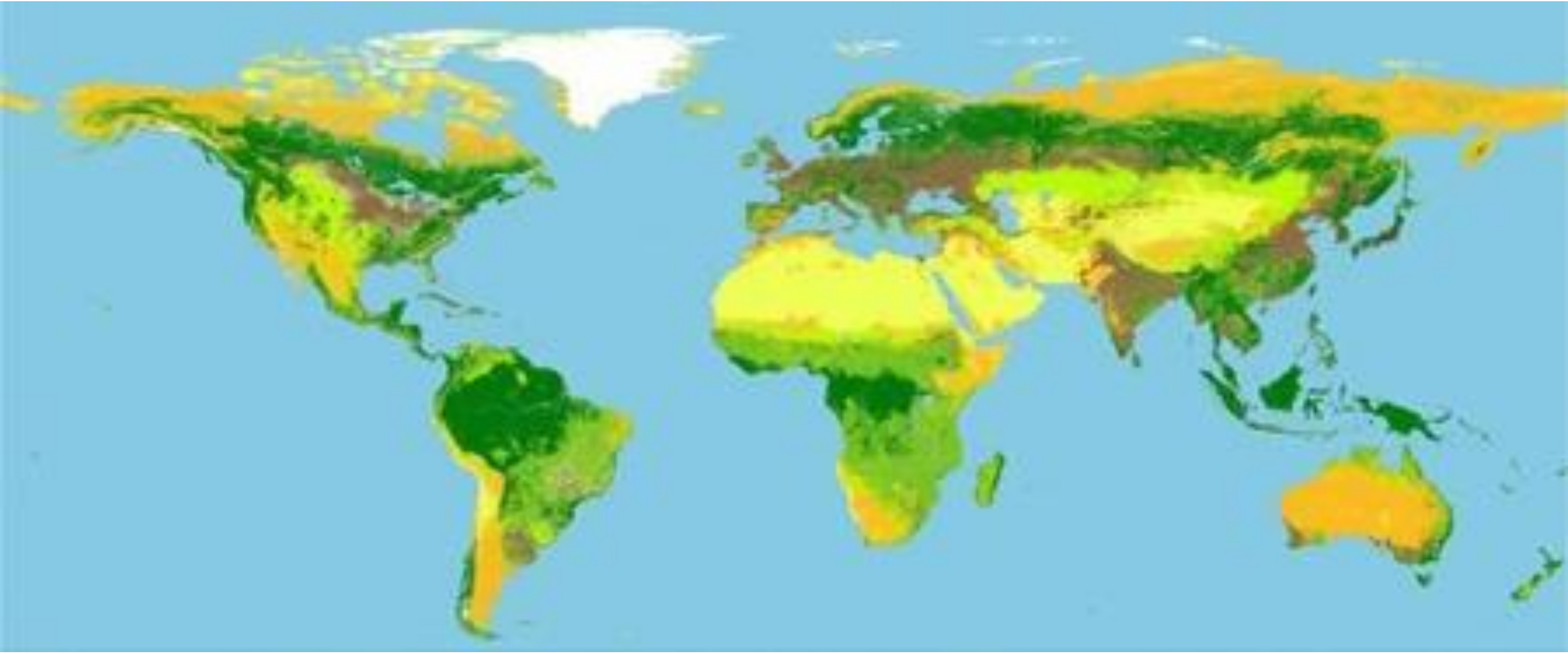
دکترای مهندسی نقشه برداری، گرایش GIS

محقق ارشد دانشگاه Duy Tan University، ویتنام

کارشناس GIS اداره کل سامانه های مکانی، سازمان نقشه برداری ایران

Email: Ahmadlou@yahoo.com

مرداد ۱۴۰۳



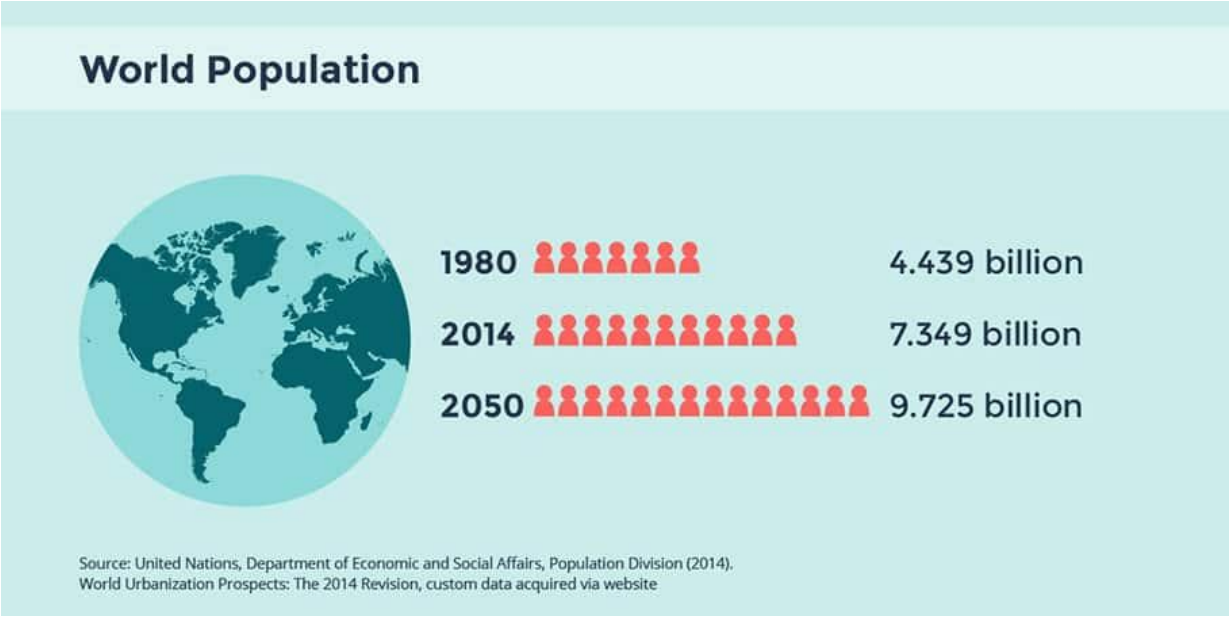
مقدمه

- داده های مورد استفاده برای مدلسازی پدیده های مختلف ماهیتی نامتوازن دارند
- مثال:
  - مدلسازی و تشخیص لکه های نفتی در بستر دریا
  - کشف تقلب از میان اسناد مختلف
  - کلاسه بندی یک تصویر ماهواره ای
  - پیش بینی بیماری های مختلف مانند آسم
  - پهنه بندی مخاطرات طبیعی در یک پهنه جغرافیایی
  - تشخیص تومورهای مغزی از تصاویر پزشکی

- عدالت در تغذیه الگوریتم های مختلف داده کاوی و یادگیری ماشین و عمیق به لحاظ تعداد نمونه در هر کلاس
- توجه قرار دادن هدف مدلسازی و تعریف مسئله
- نامتوازنی پنهان
- توجه به جمله معروف: Garbage in, garbage out
- استفاده از معیارهای ارزیابی مناسب در صورت عدم توجه به مسئله نامتوازنی

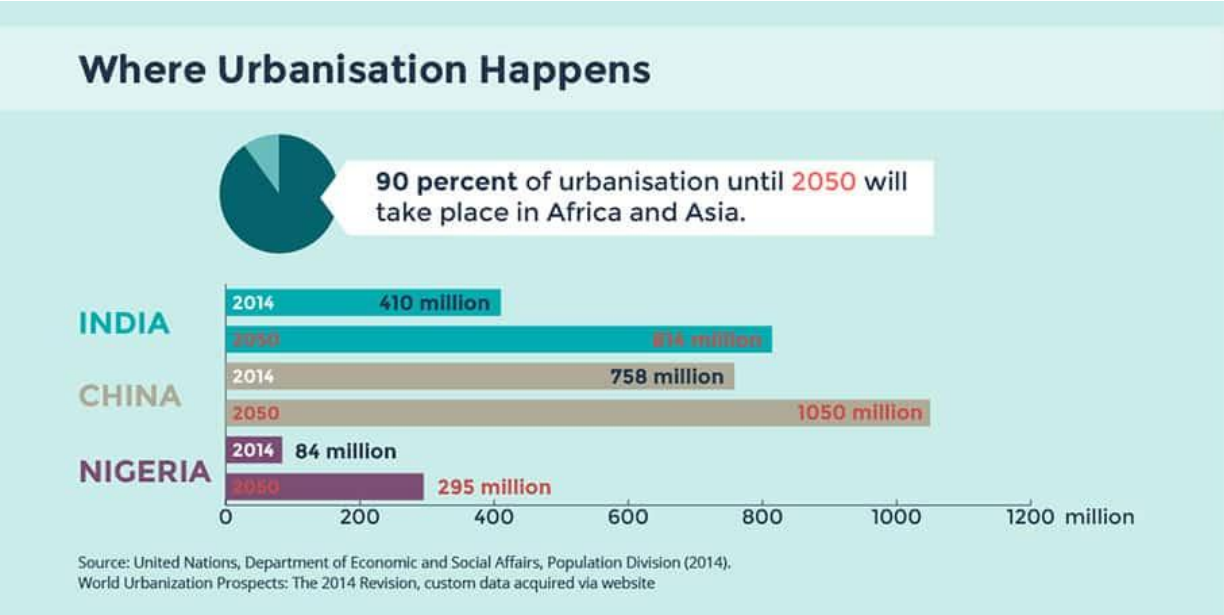
- رویکردهای موجود برای حل مسئله نامتوازنی:
  - رویکرهای سطح داده
    - کم نمونه برداری
      - نمونه برداری متوازن
    - بیش نمونه برداری
      - تولید نمونه های تکراری از کلاس اقلیت
  - رویکردهای سطح الگوریتم
    - الگوریتم های Cost-Sensitive
  - رویکردهای ترکیبی

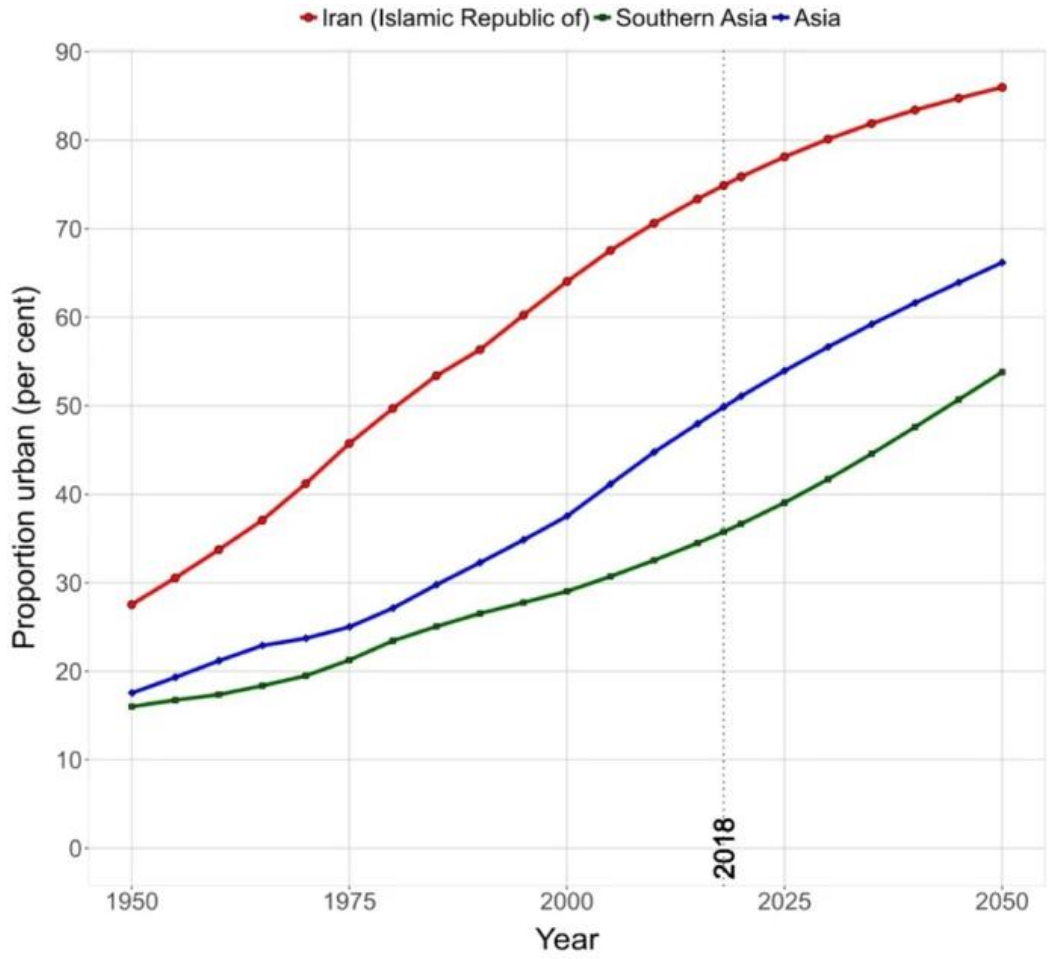
- ۱.۷ برابر شدن جمعیت کره زمین از سال ۱۹۸۰ میلادی تا سال ۲۰۱۴ میلادی
- جمعیت کنونی کره زمین: حدود ۸ میلیارد (سازمان ملل)
- پیش بینی افزایش جمعیت کره زمین به حدود ۹.۷ میلیارد نفر تا سال ۲۰۵۰



- افزایش ۱۵درصدی جمعیت شهرنشینی در سال ۲۰۱۵ نسبت به سال ۱۹۸۰
- پیش بینی ۶۶ درصدی جمعیت شهرنشینی نسبت به کل جمعیت کره زمین در سال ۲۰۵۰

- پایین بودن جمعیت شهرنشینی در قاره های آسیا و آفریقا نسبت به قاره های دیگر





Asghar Pilehvar, A. (2021). Spatial-geographical analysis of urbanization in Iran. *Humanities and Social Sciences Communications*, *8*(1), 1-12.

افزایش جمعیت شهری باعث تبدیل زمین های غیر شهری (مانند زمین های بایر، کشاورزی و جنگلی) به کاربری های شهری خواهد شد که اثرات نامطلوبی بر روی اقلیم، کیفیت آب، تراکم کشاورزی، افزایش مصرف انرژی، افزایش هزینه ایجاد زیرساخت های شهری و تنوع زیستی و جنگل زدایی خواهد شد.

پایش تغییرات کاربری اراضی شهری و مدلسازی آن می تواند اطلاعات مفیدی را درباره الگوهای رشد شهری و پارامترهای موثر در آن در اختیار برنامه ریزان شهری قرار دهد. این اطلاعات توسط برنامه ریزان و مدیران شهری برای تصمیم گیری های بهتر در حوزه های زیست محیطی، اقتصادی و اجتماعی بکار گرفته می شوند.

مدلسازی نقش اساسی در درک و برنامه ریزی برای چالش های ناشی از رشد شهری دارد که می تواند منجر به تخصیص منابع موثر محیط های شهری پایدار شود.

- تعریف مسئله
- ضرورت و انگیزه های پژوهش
- اهداف پژوهش
- سوالات پژوهش

داده های مدلسازی رشد شهری دارای نامتوازنی شدیدی بین کلاس های رشد شهری و عدم رشد شهری هستند. استفاده از مدلهای استاندارد که دارای پیشفرض برابری تعداد نمونه های کلاسها برای مدلسازی هستند برای چنین داده هایی امکان پذیر نیست.

تاکنون رویکردهای مناسب برای غلبه بر نامتوازنی بین کلاسهای رشد و عدم رشد شهری ارائه نشده است و مدلسازی های انجام گرفته با استفاده از نمونه برداری تصافی است که با توجه به نابرابری نمونه های دو کلاس رویکرد مناسبی نیست.

بنابراین، ارائه راهکارهایی برای مقابله با مسئله نامتوازنی و عدم قطعیت های ناشی از آن امری لازم و ضروری در مدلسازی رشد شهری و حتی غالب مدلسازی های مکانی دیگر است.

استفاده از نمونه برداری متوازن و انتخاب نمونه ها به تعداد مساوی از دو کلاس زمانی صحیح است که این نمونه ها دارای کمترین عدم قطعیت باشند درحالیکه با نمونه برداری تصادفی چنین امری امکان پذیر نیست

با مرور عمیق تحقیقات انجام گرفته، به وضوح مشخص است که در یکی دو سال اخیر رویکردهایی با نگاه سطحی و بدون در نظر گرفتن صحت آنها توسط محققان برای غلبه بر مسئله نامتوازنی در مدلسازی رشد شهری ارائه شده است.

هدف اصلی:

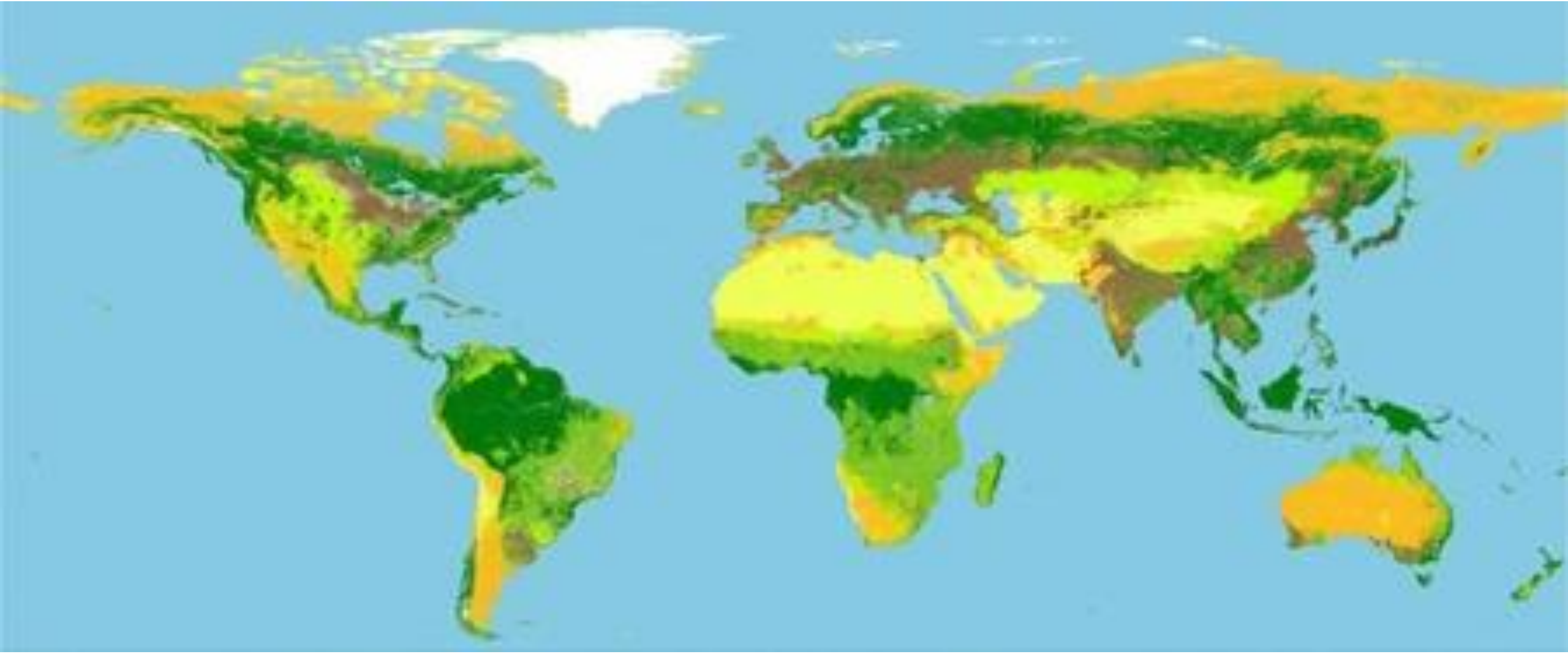
حل مسئله نامتوازنی و عدم خلوص نمونه های عدم تغییر شهری با ارائه یک منطق نمونه برداری و الگوریتم های حساس به هزینه.

- فیلتر کردن نمونه های عدم تغییر استفاده شده در مدلسازی رشد شهری با استفاده از دو رویکرد آنتروپی بیشینه و تحلیل عاملی آشیان بوم شناختی.
- انتخاب نمونه های عدم تغییر دارای بیشترین تنوع از جهت محرک های تغییرکاربری برای ورود به مجموعه داده آموزشی.
- مدلسازی رشد شهری با استفاده از مجموعه داده آموزشی تهیه شده از طریق رویکرد نمونه برداری پیشنهاد شده در این تحقیق.
- حل مسئله نامتوازنی با استفاده از الگوریتم های حساس به هزینه و مقایسه آن با مدلهای استاندارد توسعه داده شده از طریق منطق نمونه برداری ارائه شده.

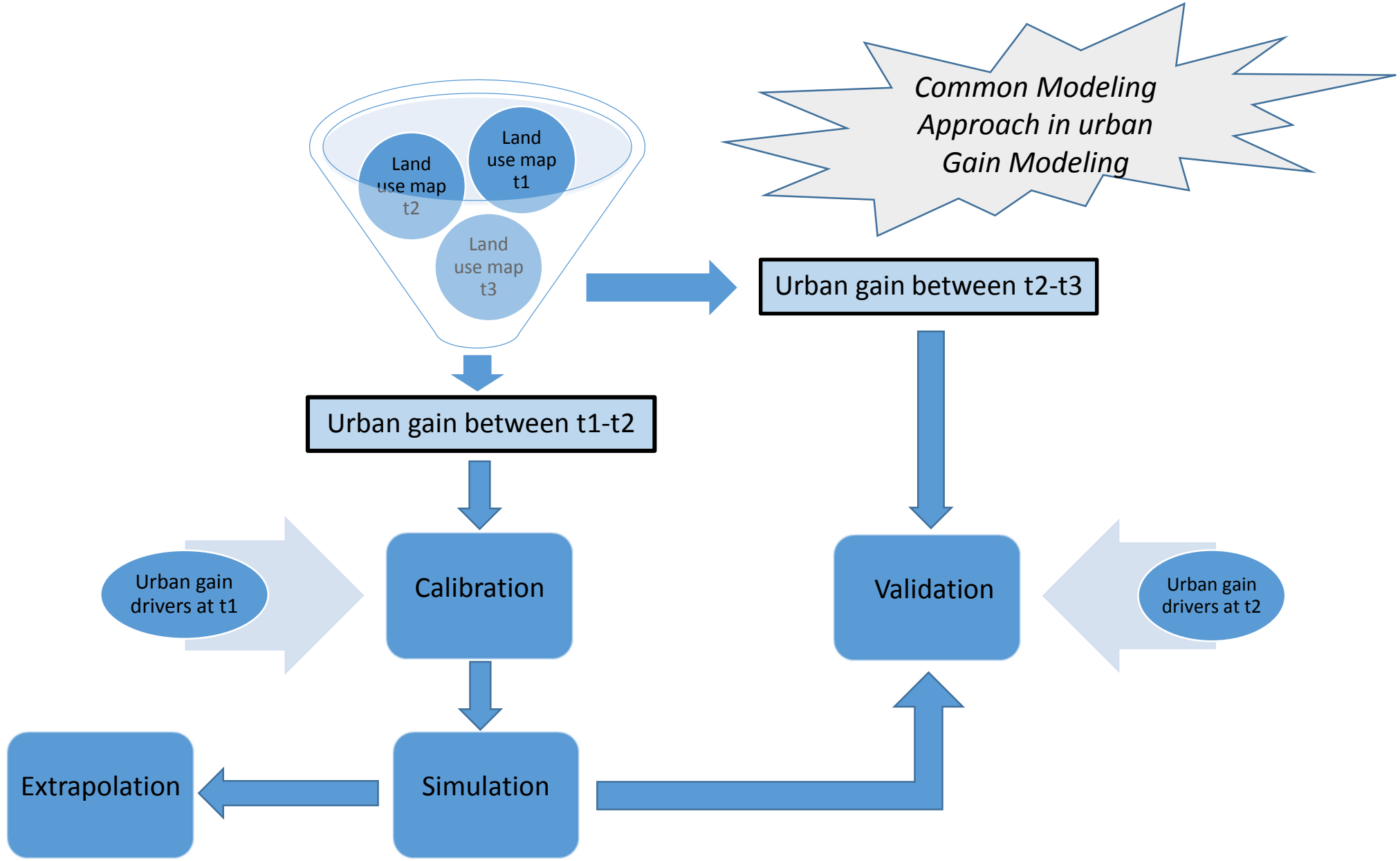
- تعریف مسئله
- ضرورت و انگیزه های پژوهش
- اهداف پژوهش
- سوالات پژوهش

سوال اصلی: در نظر نگرفتن مسئله نامتوازنی و عدم خلوص نمونه های تغییرنیافته کاربری چه تاثیری در خروجی مدل های رشد شهری دارد؟ عدم قطعیت رویکردهای نمونه برداری معمول و مدل های رایج استفاده شده در مدلسازی رشد شهری تا چه میزان در خروجی مدل ها تاثیرگذار است؟

- حذف نمونه های عدم تغییر با پتانسیل تغییر از مجموعه داده آموزشی چه تاثیری در خروجی مدل ها دارد؟
- ایجاد تنوع در میان نمونه های عدم تغییر فیلتر شده برای مجموعه داده آموزشی و انتخاب نمونه هایی با بیشترین تفاوت چه تاثیری در دقت مدل های رشد شهری دارد؟
- آیا می توان با توسعه مدل های حساس به هزینه برای مدلسازی رشد شهری از رویکردهای متداول نمونه برداری استفاده کرد؟
- کدامیک از رویکردهای سطح داده و سطح الگوریتم می توانند مسئله نامتوازنی و عدم خلوص نمونه های تغییرنیافته شهری را به خوبی حل کنند؟
- افزایش میزان نامتوازنی بین کلاس های تغییر و عدم تغییرکاربری در مدلسازی رشد شهری شهرهای اصفهان و تبریز چه تاثیری در خروجی مدل های رشد شهری دارد؟



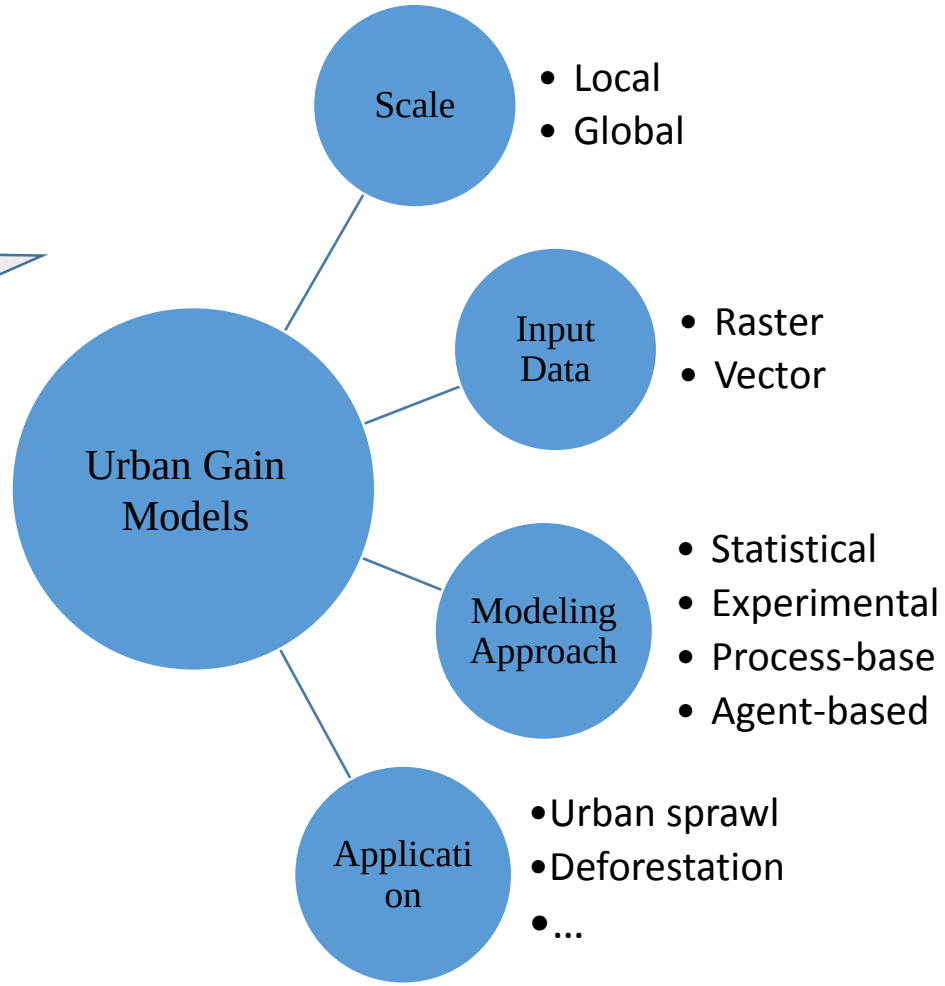
مفاهیم نظری و پیشینه پژوهش



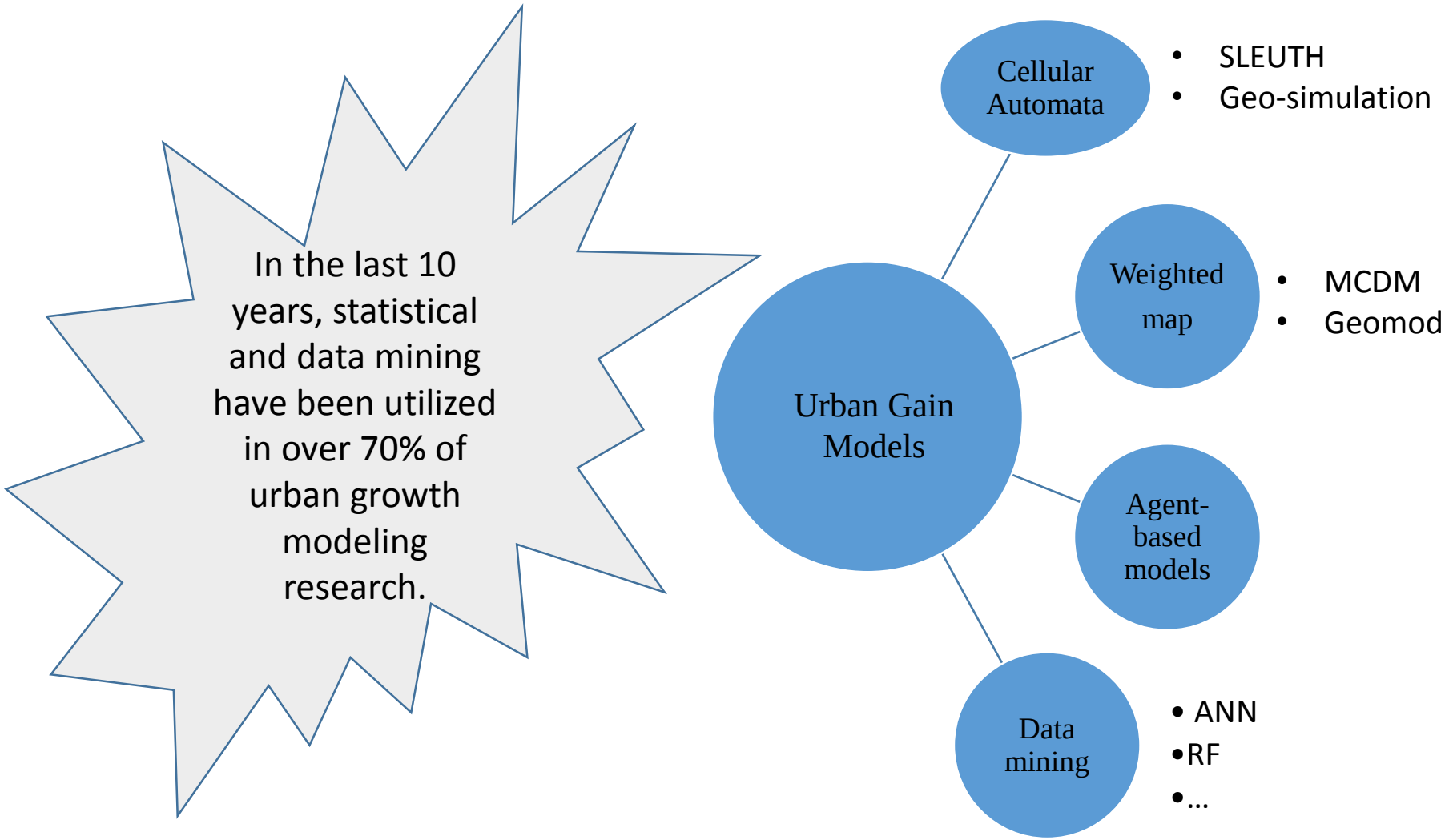
- کالبیراسیون و اعتبارسنجی
- مدل‌های رشد شهری
- تحقیقات پیشین
- جمع‌بندی

تقسیم بندی مدل های رشد شهری از دیدگاههای مختلف

Statistical and data mining models have captured much attention of researchers and modelers



تقسیم بندی مدل های رشد شهری در مطالعه حاضر



تحقیقات انجام گرفته برای مدلسازی رشد شهری شهر تبریز

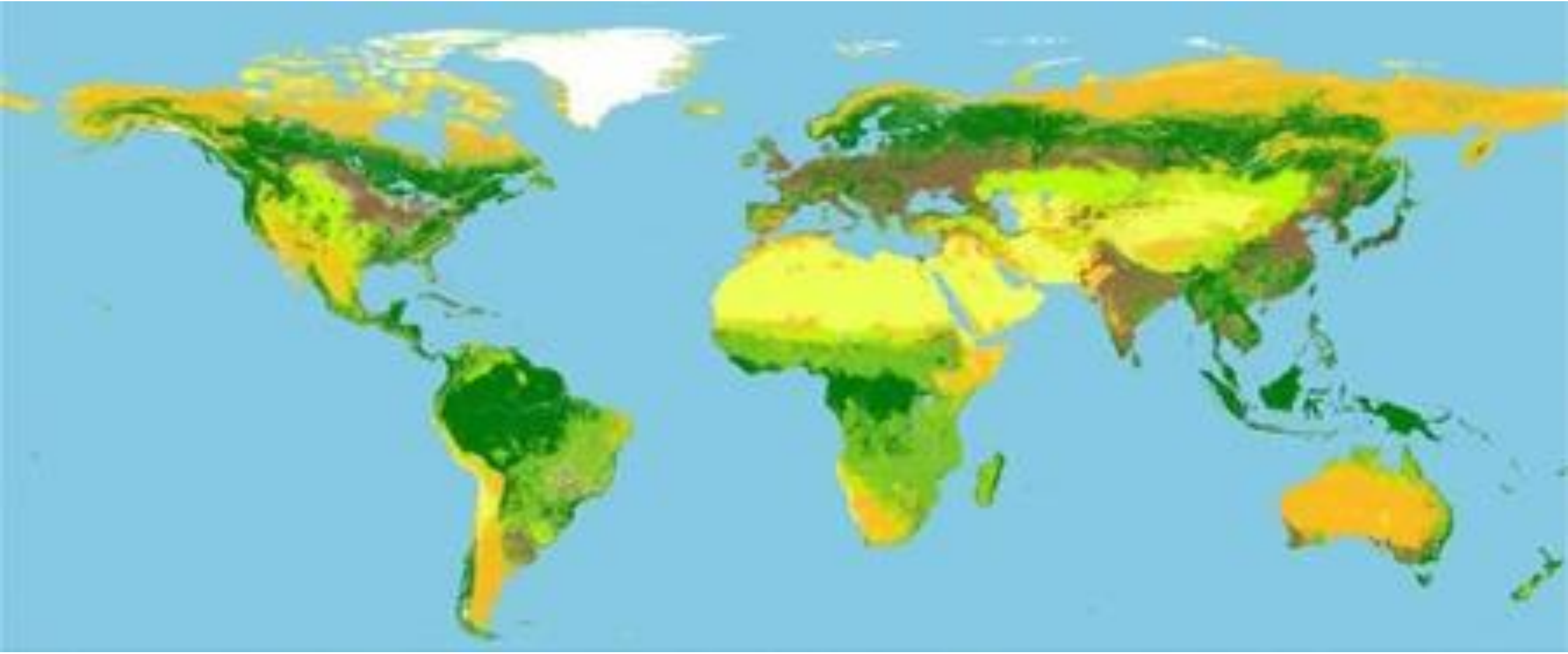
ردیف	نویسندگان (سال)	مدل استفاده شده	شاخص ارزیابی استفاده شده	نتایج	بازه‌های زمانی در نظر گرفته شده
۱	Misagh er al. (2018)	CA-Markov	کاپا دقت کلی	۸۱ درصد ۹۱ درصد	۱۳۷۲-۱۳۸۲ ۱۳۸۲-۱۳۹۲
۲	Dadashpoor et al. (2019)	SLEUTH	کاپا	دقت ۷۵ درصد	۱۳۷۴-۱۳۸۴ ۱۳۸۴-۱۳۹۴
۳	Parvinnezhad et al. (2021)	SVR-FRST	دقت کلی TPR TNR	۸۳.۶ درصد ۴۱.۶ درصد ۹۰.۴ درصد	۱۳۶۸-۱۳۷۸ ۱۳۷۸-۱۳۸۸
۴	Mahmoudzadeh et al. (2022)	شبکه عصبی مصنوعی رگرسیون منطقی	مساحت زیر منحنی	۸۲ درصد	۱۳۶۲-۱۳۷۸ ۱۳۷۸-۱۳۹۲

تحقیقات انجام گرفته برای مدلسازی رشد شهری شهر اصفهان

ردیف	نویسندگان (سال)	مدل استفاده شده	شاخص ارزیابی استفاده شده	نتایج	بازه‌های زمانی در نظر گرفته شده
1	Shafizadeh-Moghadam er al. (2017)	CA-LTM	PA دقت کلی FoM	۴۲.۶۹ درصد ۷۲.۳۱ درصد ۲۷.۱۶ درصد	۱۳۸۲-۱۳۷۲ ۱۳۹۲-۱۳۸۲
2	Mozaffar et al. (2021)	LCM	کاپا	دقت ۸۸ درصد	۱۳۸۶-۱۳۷۵ ۱۳۹۵-۱۳۸۶

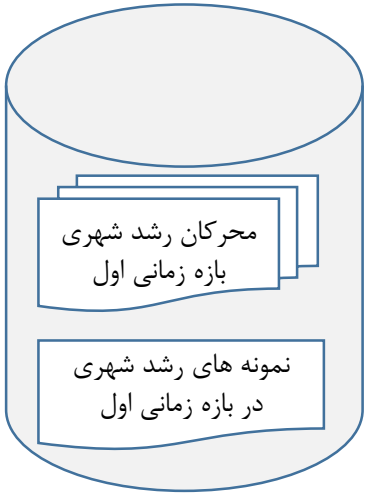
جمع بندی تحقیقات پیشین

- داده‌های استفاده شده در مدلسازی تغییرات کاربری اراضی به خصوص مدلسازی رشد شهری داده‌های نامتوازن هستند و الگوریتم‌های متداول آماری و داده کاوی نمی‌توانند برای این داده‌ها مناسب باشند.
- تاکنون در ۳۰ سال گذشته غیر از رویکرد نمونه برداری متوازن که به انتخاب نمونه‌های تغییر و عدم تغییر کاربری بصورت یکسان اقدام میکند هیچ مطالعه‌ای به ماهیت نامتوازنی داده‌های تغییرات کاربری اراضی توجه نداشته و اکثر مدلسازی به ترکیب چند مدل با هم تمرکز داشته‌اند.
- نمونه‌های عدم تغییر استفاده شده در مطالعات پیشین دارای عدم قطعیت بسیار زیاد در ارتباط با خلوص آنها می‌باشند.



روش شناسی پژوهش

ارائه یک منطق نمونه برداری با استفاده از رویکردهای سطح داده



محاسبه پتانسیل تغییر برای نمونه های عدم تغییر با استفاده از دو رویکرد آنتروپی بیشینه و تحلیل عامل آشیان بوم شناختی

حذف نمونه های عدم تغییری که مقدار آنتروپی بیشینه و تحلیل عاملی آنها بالاتر از حد آستانه ۰.۵ است

استخراج نمونه های عدم تغییر باقی مانده

رویکرد سطح داده

رویکرد سطح الگوریتم

مدل‌ها

خوشه بندی نمونه های عدم تغییر با استفاده از الگوریتم K-means

اعتبارسنجی خوشه بندی با استفاده از شاخص های دیویس بولدین و شاخص هاراباز

انتخاب نمونه های عدم تغییر به تعداد نمونه های تغییریافته بازه زمانی اول به تعداد مساوی از هر خوشه

تکرار نمونه برداری با استفاده از نمونه برداری مجدد Bootstrap برای نمونه های عدم تغییر

تشکیل دیتاست های آموزشی از طریق داده های بازه زمانی اول برای ایجاد مدلهای آماری و یادگیری ماشین

استخراج نمونه های عدم تغییر باقی مانده

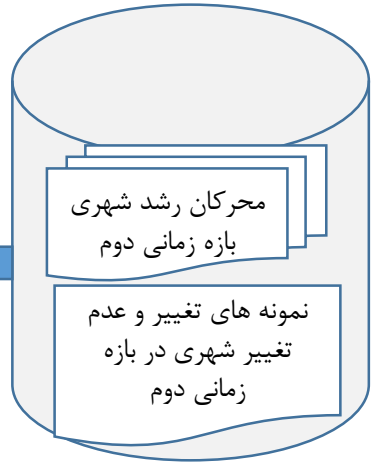
مدلسازی رشد شهری با استفاده از الگوریتم های داده کاوی مانند شبکه عصبی مصنوعی، جنگل تصادفی و ماشین بردار پشتیبان

ساخت مدلهای تک کلاسه آنتروپی بیشینه و تحلیل عاملی

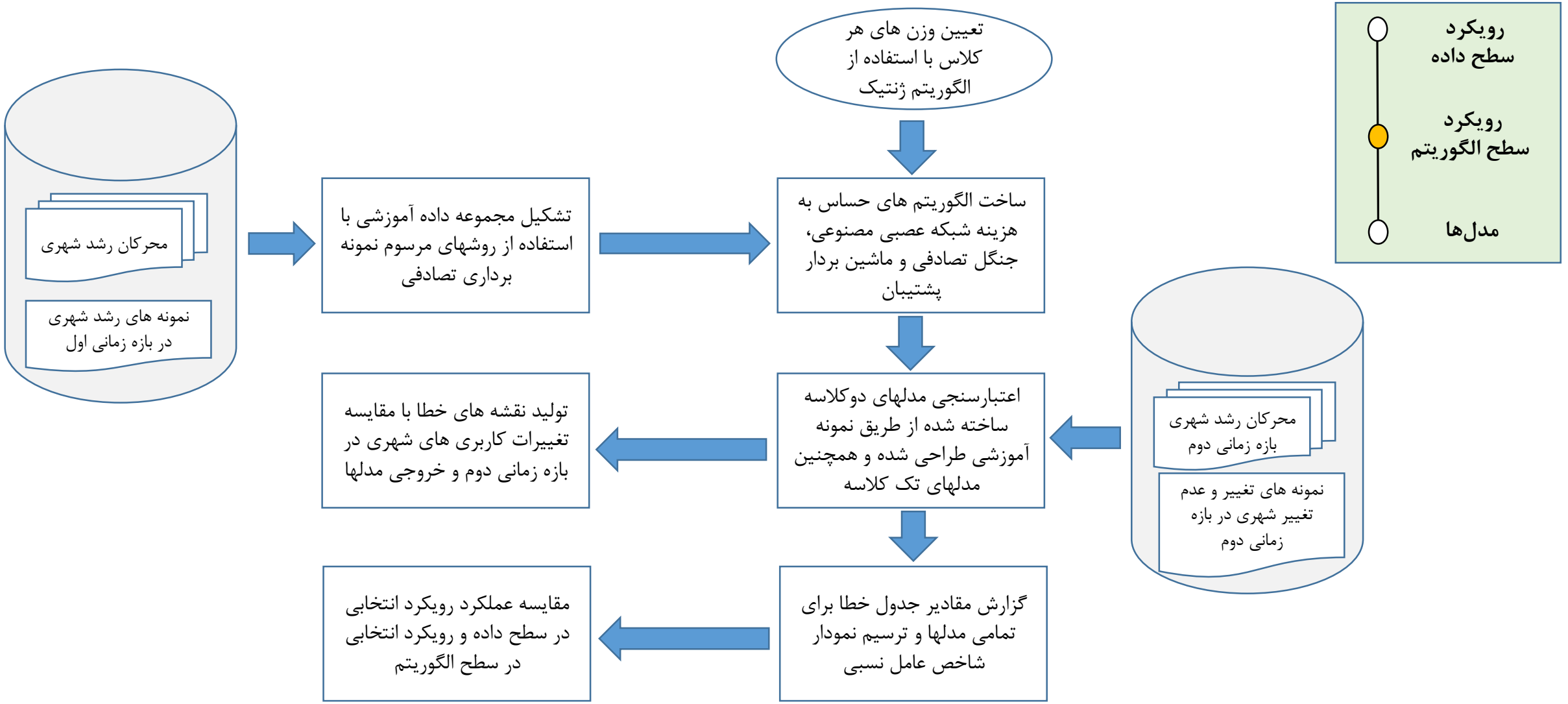
تولید نقشه های خطا

اعتبارسنجی مدلهای دوکلاسه ساخته شده از طریق نمونه آموزشی طراحی شده و همچنین مدلهای تک کلاسه

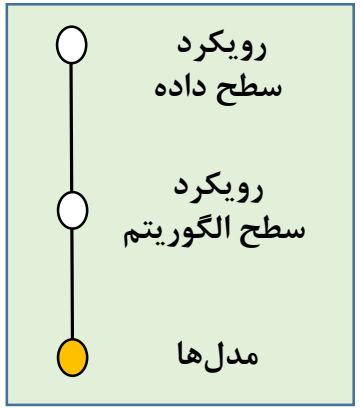
خروجی گرفتن مقادیر جدول خطا و شاخص عامل کلی



مقابله با مسئله نامتوازنی با ارائه راهکاری در سطح الگوریتم با استفاده از مدلهای حساس به هزینه



آنتروپی بیشینه



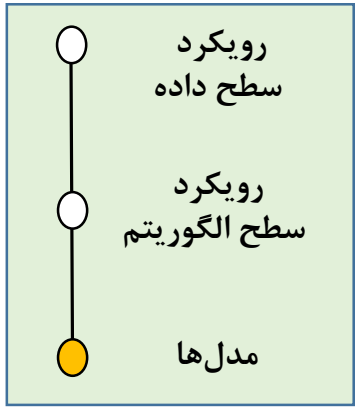
$$H(\bar{P}) = -\sum_{x \in X} \bar{P}(x) ln\, \bar{P}(x)$$

در این رابطه ln لگاریتم طبیعی است. بر اساس مسئله دوگان محدب (Convex Duality) و با به حداکثر رساندن آنتروپی با توجه به قیدهای مسئله، توزیع گیبس (Gibbs Distribution) تنها توزیعی است که کوچکترین Kullback-Leibler را دارا بوده و تمام قیدها را بدون پیشفرض‌های اضافی ارضا می‌کند.

$$\bar{P}\,(X) = \frac{1}{n}\sum_{i=1}^{n} f_j\,(x_i)$$

$\bar{P}\,(X)$ میانگین ویژگیها یا محرکها تحت توزیع

مدل تحلیل عاملی آشیان بوم شناختی



- مدلی چند متغیره که از دو مفهوم تحلیل فاکتور (Factor Analysis) و نظریه آشیان اکولوژیکی (Ecological Niche Theory) برای مطالعه توزیع گونه‌ها (Species) بر اساس متغیرهای پیش‌بینی‌کننده و مکان‌های صرفا حضور گونه‌ها و بدون نیاز به مکان‌های عدم حضور ارائه شده است

- مدل تحلیل فاکتور تحلیل عاملی آشیان بوم شناختی تفاوت بین توزیع متغیرهای پیش بینی شده در سلول های تغییریافته را نسبت به سایر سلول ها را برای تمام منطقه مورد مطالعه محاسبه می کند.

- در مطالعه حاضر سلول های توسعه یافته شهری معادل یک گونه تلقی شده اند که در طول زمان در مناطق مختلف توسعه پیدا کرده اند.

خوشه بندی K-means و شاخص های دیویس بولدین و هاراباز

رویکرد سطح داده

رویکرد سطح الگوریتم

مدل‌ها

- بخاطر پایه ای بودن منطق نمونه برداری طراحی شده، روش خوشه بندی K-means که یکی از ساده ترین و مشهورترین الگوریتم های خوشه بندی است برای این منظور انتخاب شد.
- پیاده سازی این الگوریتم آسان بوده و مناسب برای داده های با حجم بالا می باشد.

گام اول: در ابتدا از بین n نقطه داده‌ای، K نقطه $z_1, z_2, \ldots, z_k$ به عنوان مراکز خوشه‌ها انتخاب می‌شوند.

گام دوم: نقاط داده‌ای x_i $(i=1, 2, \ldots, n)$ را به خوشه‌ی فرضی C_j تخصیص داده می‌شود اگر و تنها اگر رابطه زیر برقرار باشد:

$$\|x_i - z_j\| < \|x_i - z_p\| \quad , \quad P = 1, 2, 3, \ldots, K \quad , j \neq p$$

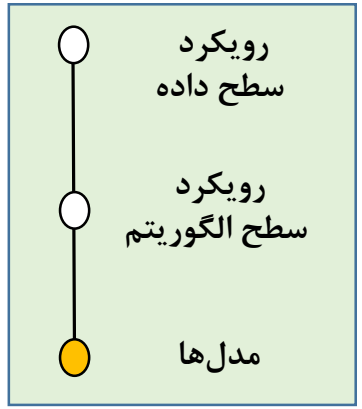
گام سوم: وقتی که تمام نمونه‌ها به خوشه‌ها تخصیص داده شدند، موقعیت k مرکز خوشه دوباره محاسبه می‌شود. مراکز جدید $z_1, z_2, \ldots, z_k$ از رابطه زیر محاسبه می‌شوند:

$$Z_i = \frac{\sum x_i}{n_i} \quad , \quad i = 1, 2, \ldots, k \quad x_j \in C_i$$

که n_i تعداد عناصر متعلق به خوشه C_i می‌باشد.

گام چهارم: گام‌های دوم و سوم آنقدر تکرار می‌شوند تا مراکز دیگر جابجا نشوند.

خوشه بندی K-means و شاخص های هارابازو دیویس بولدین

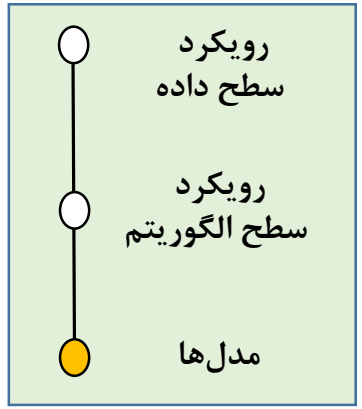


- برای پیدا کردن تعداد خوشه های بهینه در این روش از دو شاخص هارابازو دیویس-بولدین استفاده شد.
- شاخص هارابباز:

$$VRC_k = \frac{SS_B}{SS_W} \cdot \frac{N-K}{K-1}$$

این شاخص با عنوان Variance ratio criterion نیز شناخته میشود در رابطه بالا SSB مجموع واریانس بین خوشه ها و SSW مجموع واریانس داخل خوشه ها هستند. K و N به ترتیب تعداد مشاهدات و خوشه ها را نشان می دهند.

خوشه بندی K-means و شاخص های دان و دیویس بولدین



- شاخص اعتبارسنجی دیویس- بولدین:

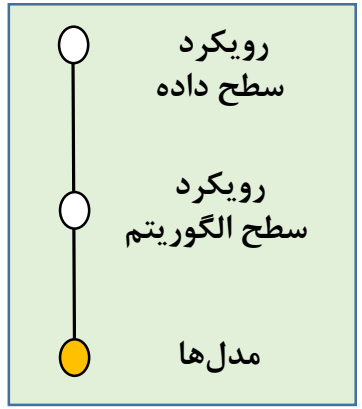
تابعی از پراکندگی نمونه های داخل هر خوشه و همچنین فاصله بین خوشه ها می باشد

$$R_{ij} = \frac{S_i + S_j}{d_{ij}}$$

در این رابطه S شاخص پراکندگی و d فاصله بین دو خوشه می باشد

$$S_i = \frac{1}{C_i} \sum_{x \in C_i} \{\|x - z_i\|\} \, , d_{ij} = \|z_i - z_j\|$$

جنگل تصادفی مبتنی بر هزینه



- رابطه شاخص جینی استاندارد:

$$Gini_c = 1 - \sum_{i=1}^{k} \left(\frac{n_i}{\sum_{i=1}^{k} n_i}\right)^2 \quad , \qquad c \in \{1, 2, \dots, H\}$$

در این رابطه c نشان‌دهنده نودهای فرزند مربوطه، k نشان دهنده تعداد کلاس‌های مورد نظر، n تعداد مشاهدات مربوط به کلاس iام و H تعداد نودهای فرزند حاصل از انشعاب هر نود می‌باشند

- برای دخیل کردن وزن‌های مربوط در رابطه بالا روابط زیر را داریم:

$$Impurity = \sum_{c=1}^{H} \frac{t_c}{t_p}\, i_c$$

$$t_c = \sum_{i=1}^{k} w_i\, n_i \ , \ i \in \{1, 2, \dots, k\}$$

$$t_p = \sum_{c=1}^{H} t_c \quad , \qquad c \in \{1, 2, \dots, H\}$$

$$i_c = 1 - \sum_{i=1}^{k} \left(\frac{w_i n_i}{t_c}\right)^2$$

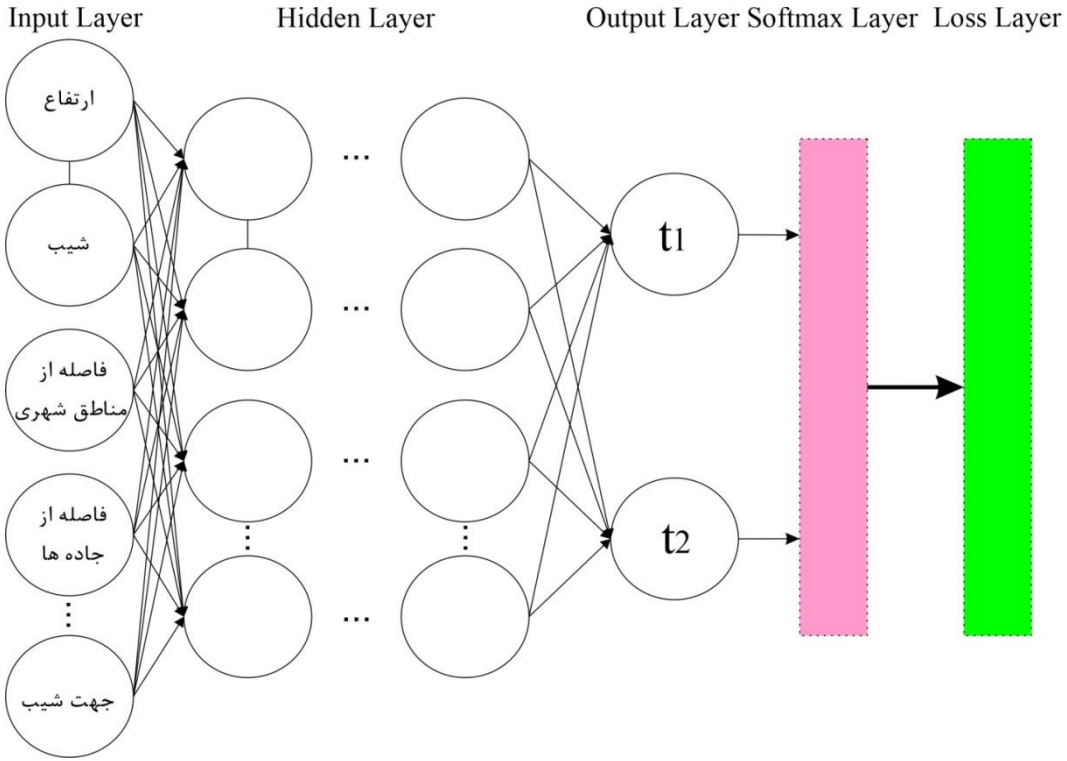
در این روابط Impurity بیانگر ناخالصی کلی می‌باشد و هدف مسئله انتخاب کوچکترین ناخالصی در حل مسئله مدلسازی می‌باشد

شبکه عصبی مبتنی بر هزینه

رویکرد سطح داده

رویکرد سطح الگوریتم

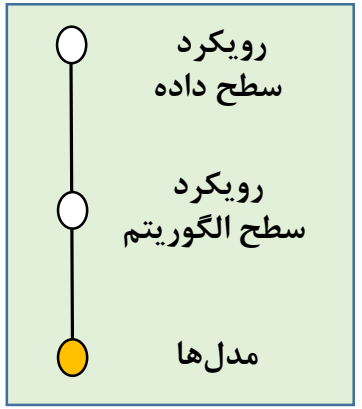
مدل‌ها



$$\vartheta'_x = -\left[y_+ \log\left(\theta_1\ (F_{(t_1)}, F_{(t_2)})\ F_{(t_1)}\right) + y_- \log\left(\theta_2\ ((F_{(t_1)}, F_{(t_2)})F_{(t_2)}\right)\right]$$

$$\theta_1\left(F_{(t_1)}, F_{(t_2)}\right) = \frac{a^{F(t_1)-\varepsilon}}{F(t_1)\ a^{F(t_1)-\varepsilon} + F(t_2)\ b^{F(t_2)-\delta}} \qquad \theta_2\left(F_{(t_1)}, F_{(t_2)}\right) = \frac{b^{F(t_2)-\delta}}{F(t_1)\ a^{F(t_1)-\varepsilon} + F(t_2)\ b^{F(t_2)-\delta}}$$

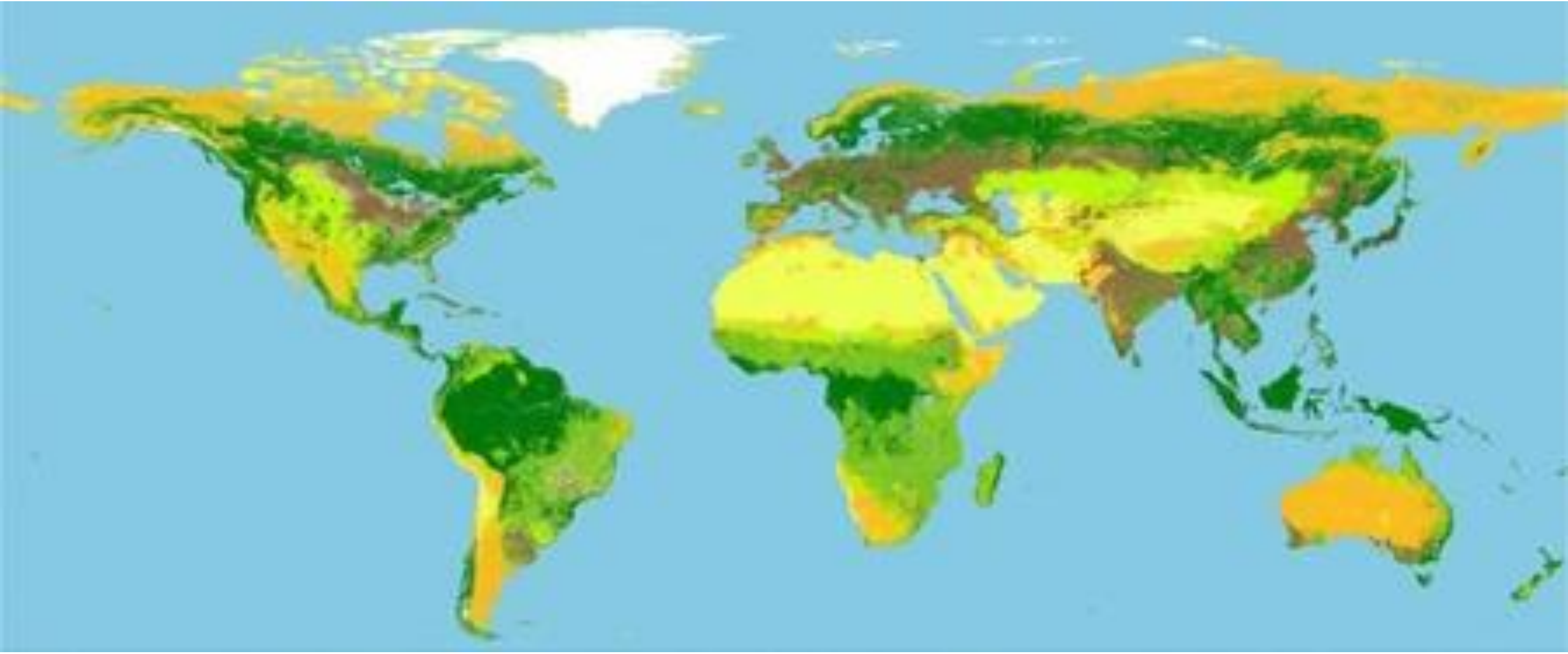
ماشین بردار پشتیبان مبتنی بر هزینه



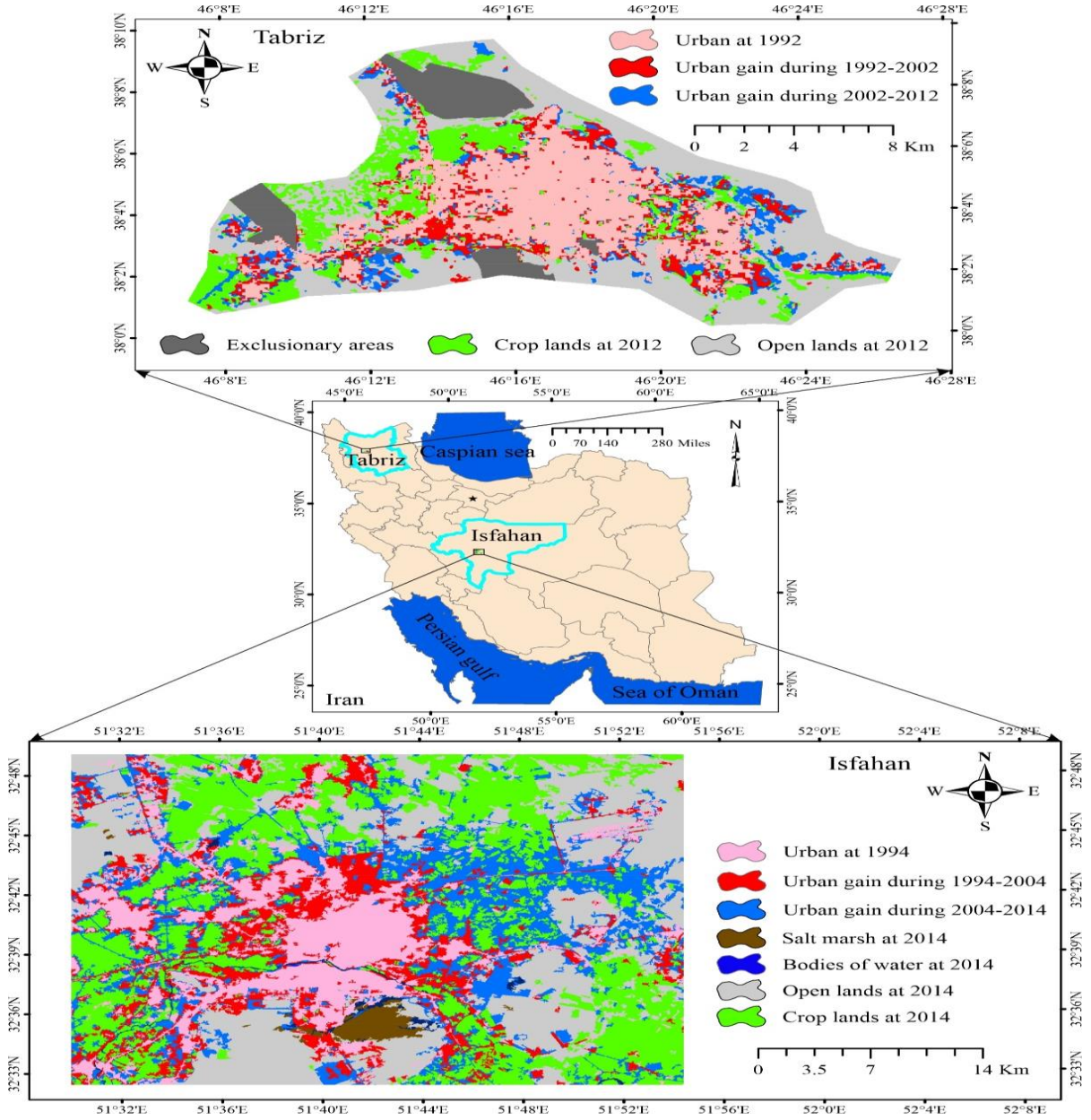
- هنگامی که داده های مورد مطالعه بصورت نامتوازن بین دو کلاس مدلسازی توزیع شده باشند مسئله بهینه سازی مدل ماشین بردار پشتیبان تبدیل به مسئله زیر می شود:

$$\min \frac{1}{2} \|w\|^2 + C_+ \sum_{i=1}^{m_1} \xi_i^+ + C_- \sum_{i=1}^{m_2} \xi_i^-$$

$$\begin{aligned} s.t. \quad & \left(w^T x^{(n)} + w_0\right) \geq 1 - \xi_i^+ , && i = 1, 2, \ldots, m_1 \\ & \left(w^T x^{(n)} + w_0\right) \leq -1 + \xi_i^- , && i = 1, 2, \ldots, m_2 \\ & \xi_i^+ \geq 0, && i = 1, 2, \ldots, m_1 \\ & \xi_i^- \geq 0, && i = 1, 2, \ldots, m_2 \end{aligned}$$



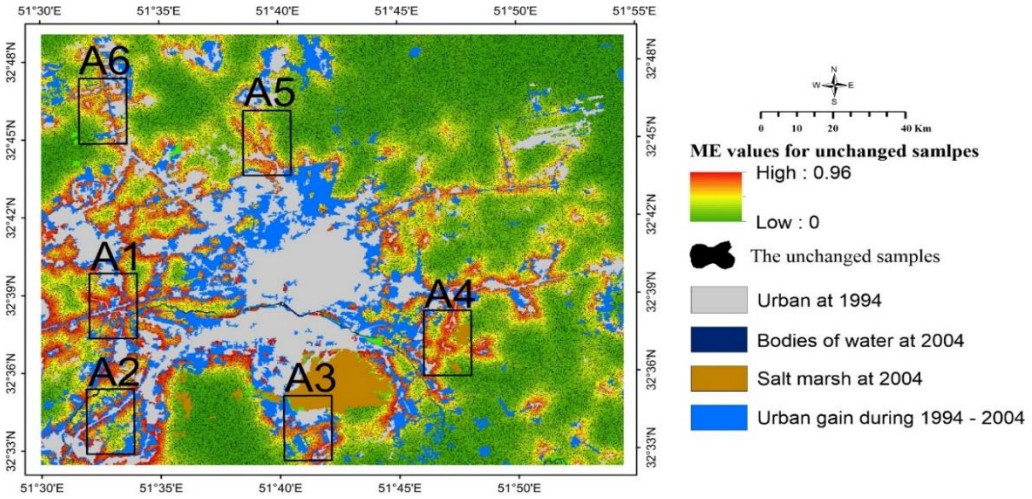
نتایج و یافته های پژوهش

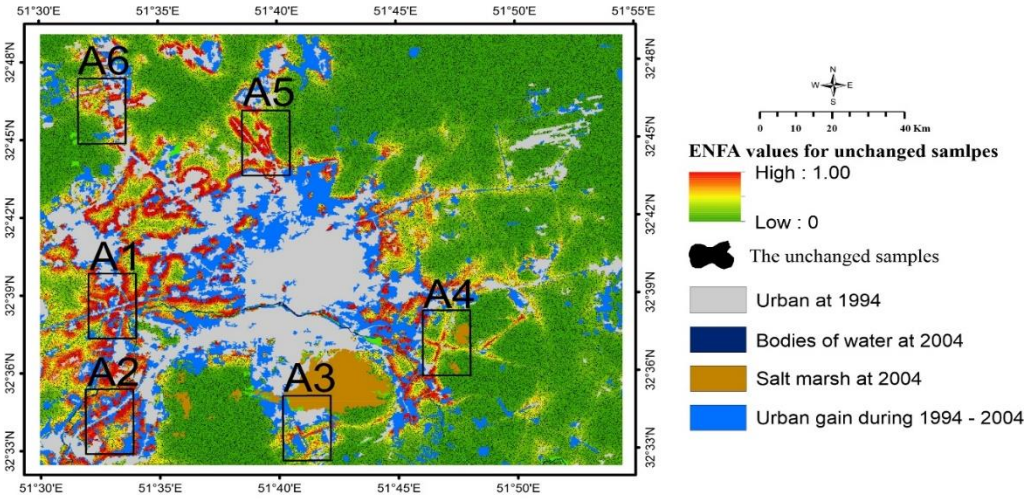


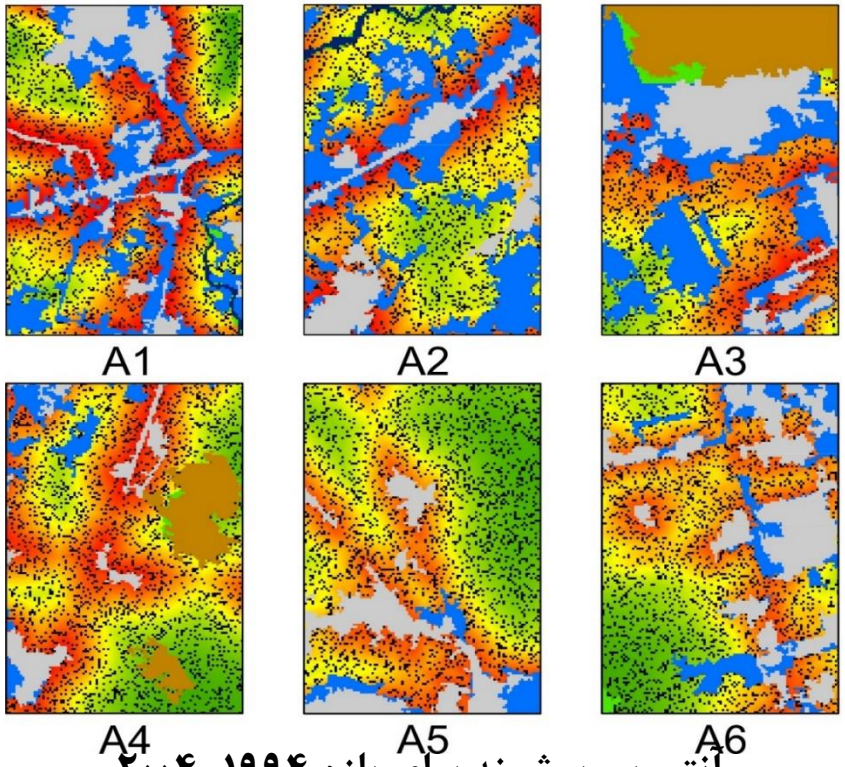
- مناطق مورد مطالعه و داده ها
- نتایج منطق نمونه برداری
- مدلسازی دوکلاسه مبتنی بر منطق نمونه برداری
- مدلسازی تک کلاسه
- اعتبارسنجی نتایج
- نتایج مدلسازی مبتنی بر هزینه
- اعتبارسنجی نتایج

فاکتورهای محرک در نظر گرفته شده برای رشد شهری

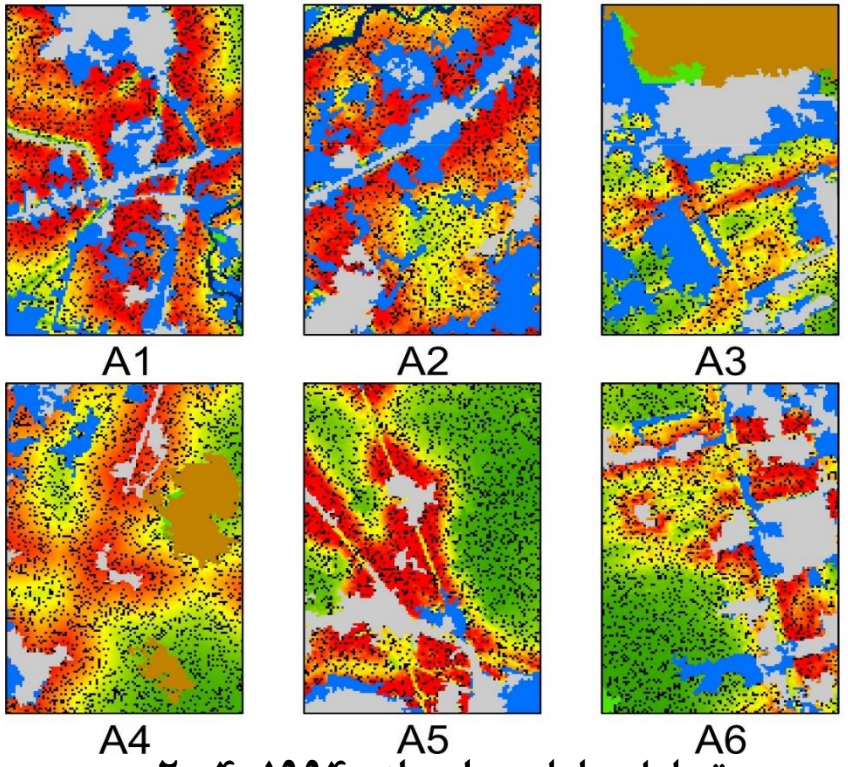
شماره ردیف	تبریز (بین سال‌های ۱۹۹۲ تا ۲۰۰۲)	اصفهان (بین سالهای ۱۹۹۴ تا ۲۰۰۴)
۱	ارتفاع	ارتفاع
۲	شیب	شیب
۳	جهت شیب	فاصله از جاده‌های اصلی
۴	فاصله از مناطق کشاورزی	فاصله از مناطق کشاورزی
۵	فاصله از مراکز تجارتی	فاصله از شوره‌زار
۶	فاصله از مناطق ساخته شده	فاصله از مناطق ساخته شده
۷	فاصله از مناطق بایر	فاصله از مناطق بایر
۸	پارامتر X	پارامتر X
۹	پارامتر Y	پارامتر Y
۱۰	فاصله از جاده‌های اصلی	–
۱۱	فاصله از گسل	–
۱۲	فاصله از شهرک‌های صنعتی	–
۱۳	فاصله از راه آهن	–
۱۴	فاصله از رودخانه	–





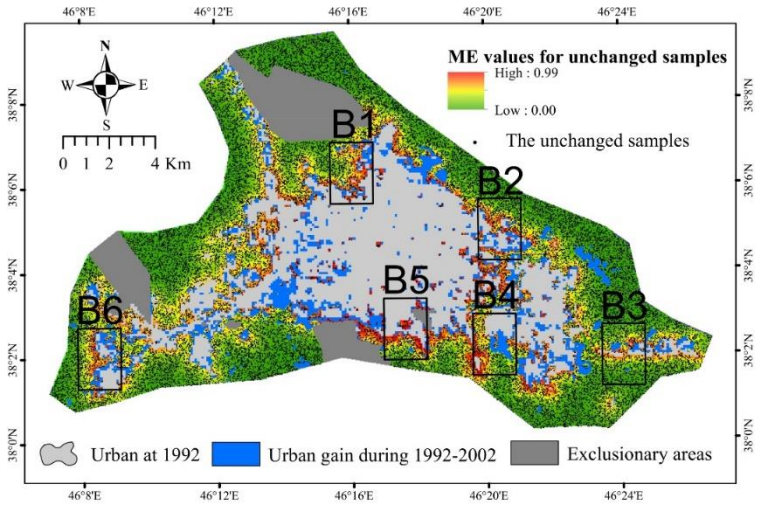


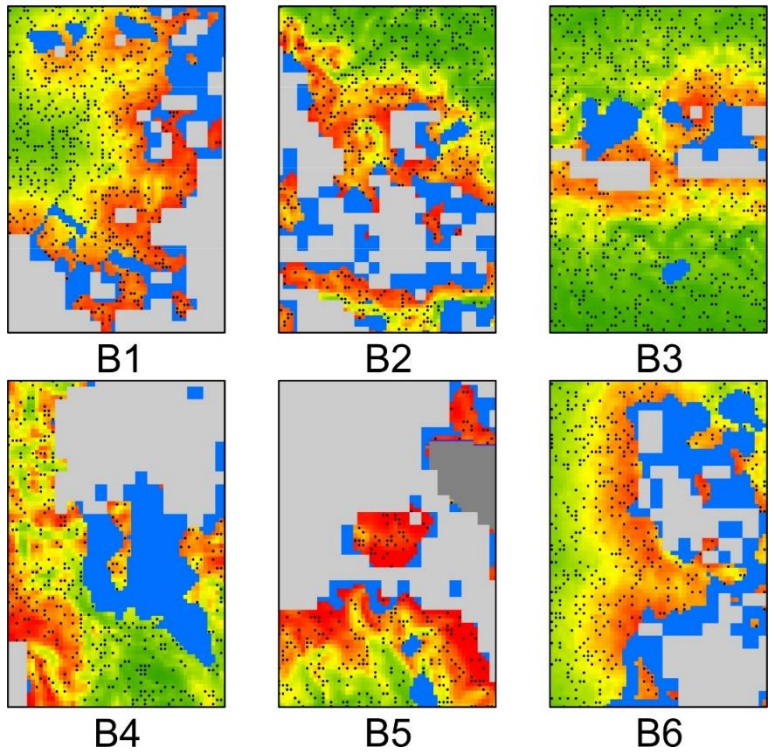
آنتروپی بیشینه برای بازه ۱۹۹۴-۲۰۰۴



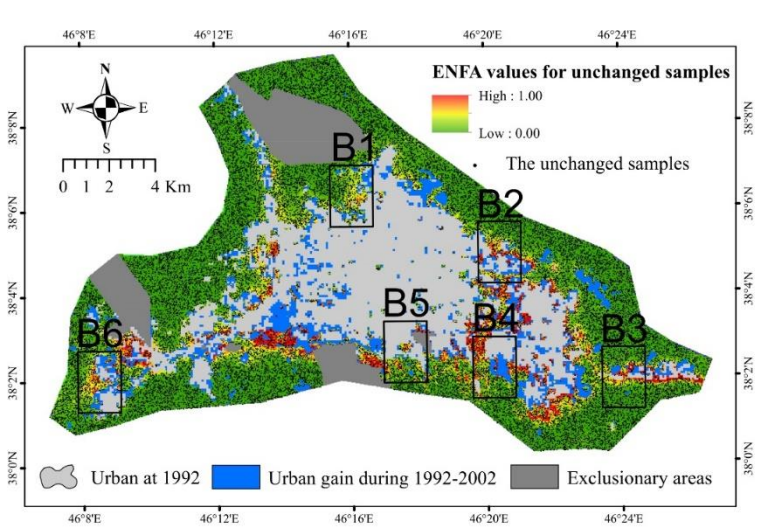
تحلیل عاملی برای بازه ۱۹۹۴-۲۰۰۴

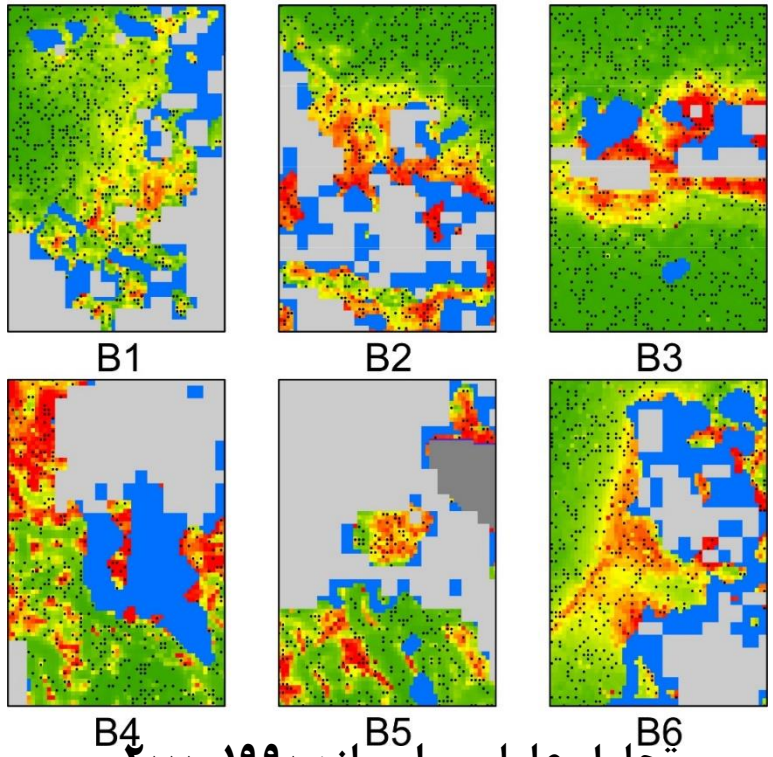
- مناطق مورد مطالعه و داده ها
- نتایج منطق نمونه برداری
- مدلسازی دوکلاسه مبتنی بر منطق نمونه برداری
- مدلسازی تک کلاسه
- اعتبارسنجی نتایج
- نتایج مدلسازی مبتنی بر هزینه
- اعتبارسنجی نتایج





آنتروپی بیشینه برای بازه ۱۹۹۰-۲۰۰۰

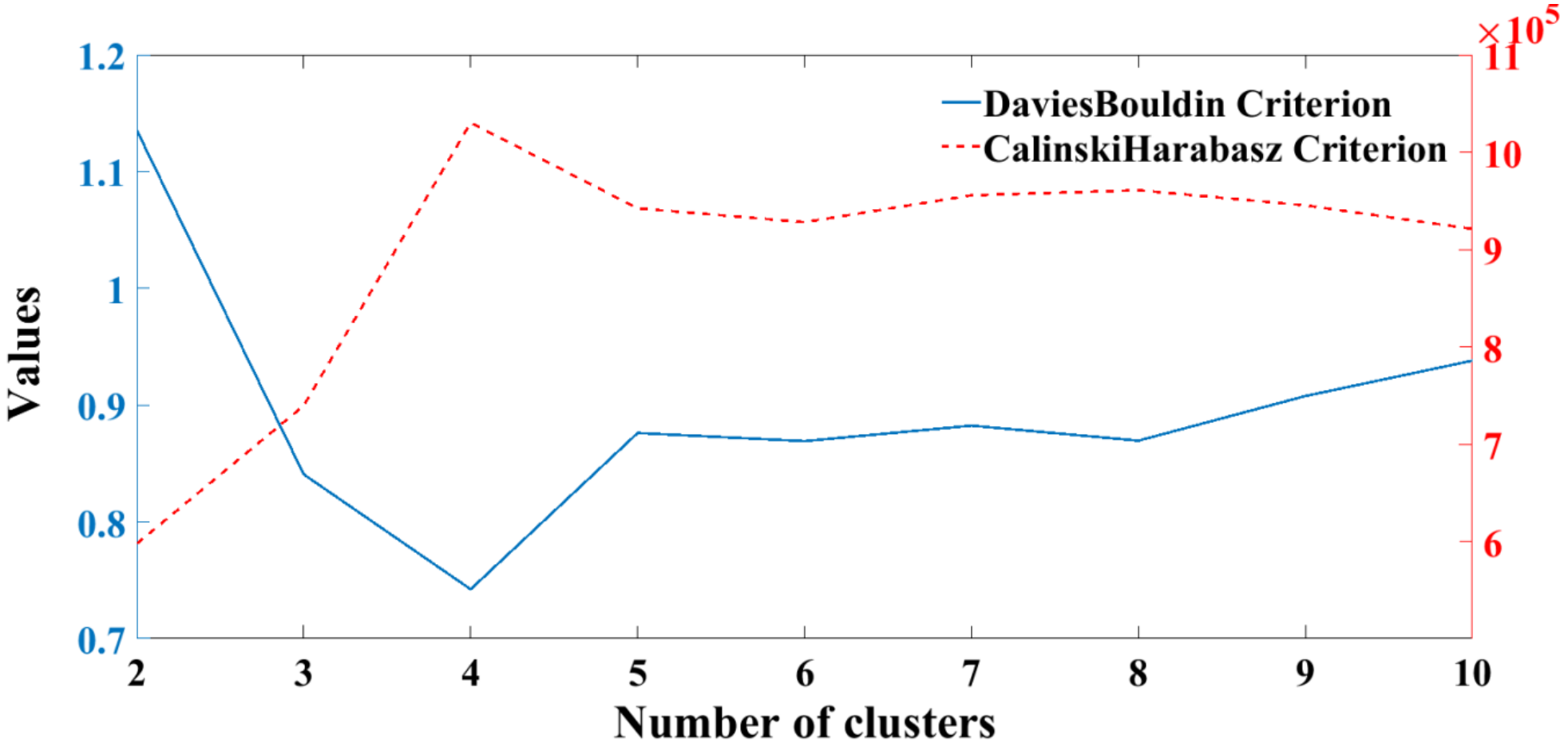




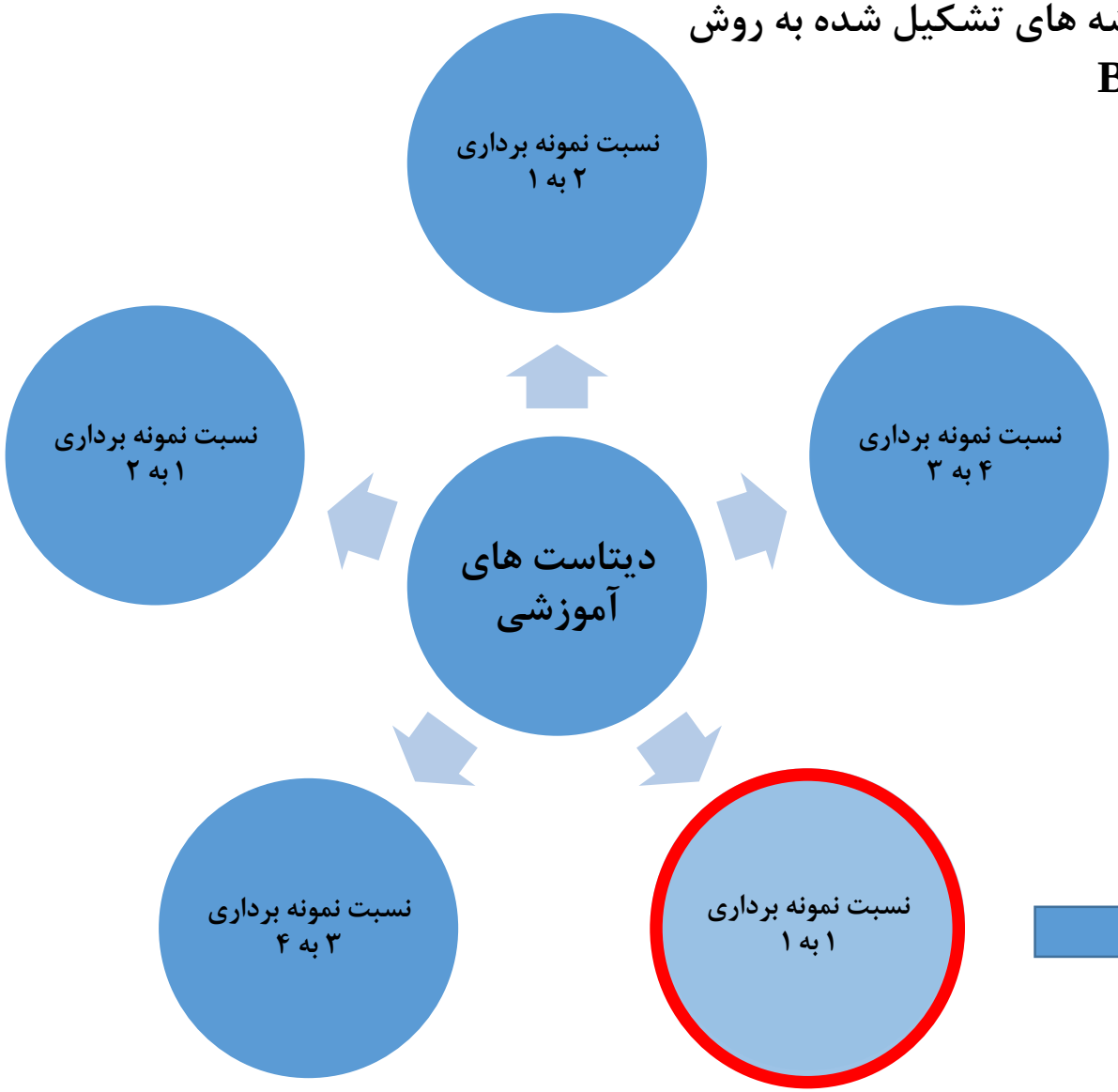
تحلیل عاملی برای بازه ۱۹۹۰-۲۰۰۰

- مناطق مورد مطالعه و داده ها
- نتایج منطق نمونه برداری
- مدلسازی دوکلاسه مبتنی بر منطق نمونه برداری
- مدلسازی تک کلاسه
- اعتبارسنجی نتایج
- نتایج مدلسازی مبتنی بر هزینه
- اعتبارسنجی نتایج

اعتبارسنجی خوشه بندی با استفاده از دو شاخص دیویس-بولدین و شاخص هارا‌باز



تشکل مجموعه داده های آموزشی متشکل از نمونه های رشد شهری بازه زمانی اول و نمونه برداری مجدد از خوشه های تشکیل شده به روش Bootstrap



GEOCARTO INTERNATIONAL
2022, VOL. 37, NO. 19, 5669–5692
https://doi.org/10.1080/10106049.2021.1923826

Taylor & Francis
Taylor & Francis Group

Check for updates

A new framework to deal with the class imbalance problem in urban gain modeling based on clustering and ensemble models

Mohammad Ahmadlou[a], Mohammad Karimi[a] and Robert Gilmore Pontius Jr[b]

[a]GIS Department, Geodesy and Geomatics Faculty, K.N. Toosi University of Technology, Tehran, Iran; [b]School of Geography, Clark University, Worcester, Massachusetts, USA

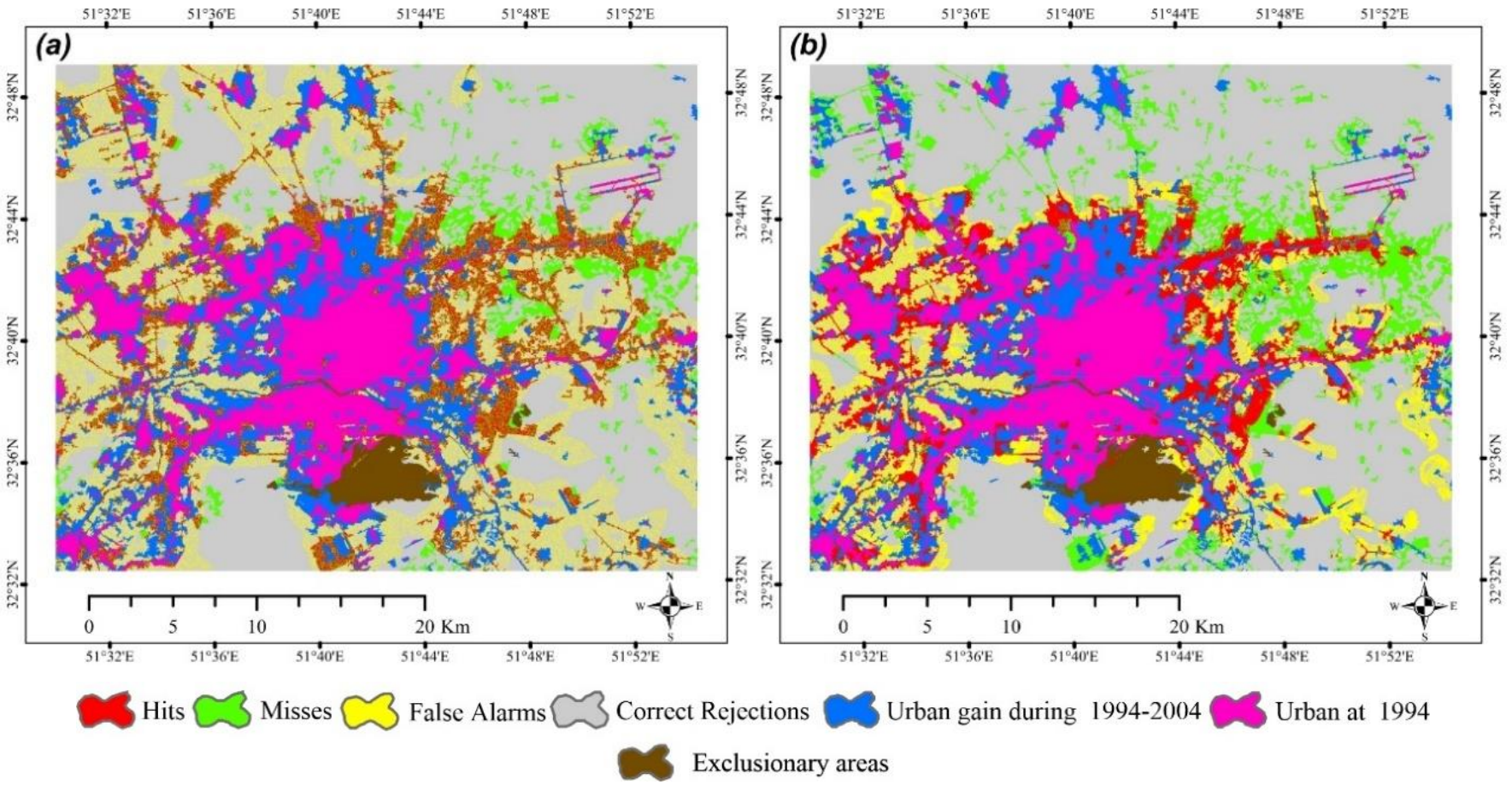
ABSTRACT

The data employed in urban gain modeling classes are often imbalanced, negatively affecting the accuracy of traditional and standard data mining and machine learning models. This study presents a new framework on the basis of clustering-based modeling and ensemble models to deal with the class imbalance problem in urban gain modeling. The random forest (RF), artificial neural network (ANN) and support vector machine (SVM) models served as the base models for the generation and evaluation of the results within this framework. The changes in urban land-use pattern of Isfahan in Iran in two time intervals of 1994-2004 and 2004-2014 were considered for the modeling. The findings showed that the proposed sampling strategy yields higher Hits and Correct Rejections rates than the strategies applied in previous studies in all three models. In the second part of the proposed framework (ensemble models), there was no substantial difference in the confusion matrix entries.

ARTICLE HISTORY
Received 17 January 2021
Accepted 5 April 2021

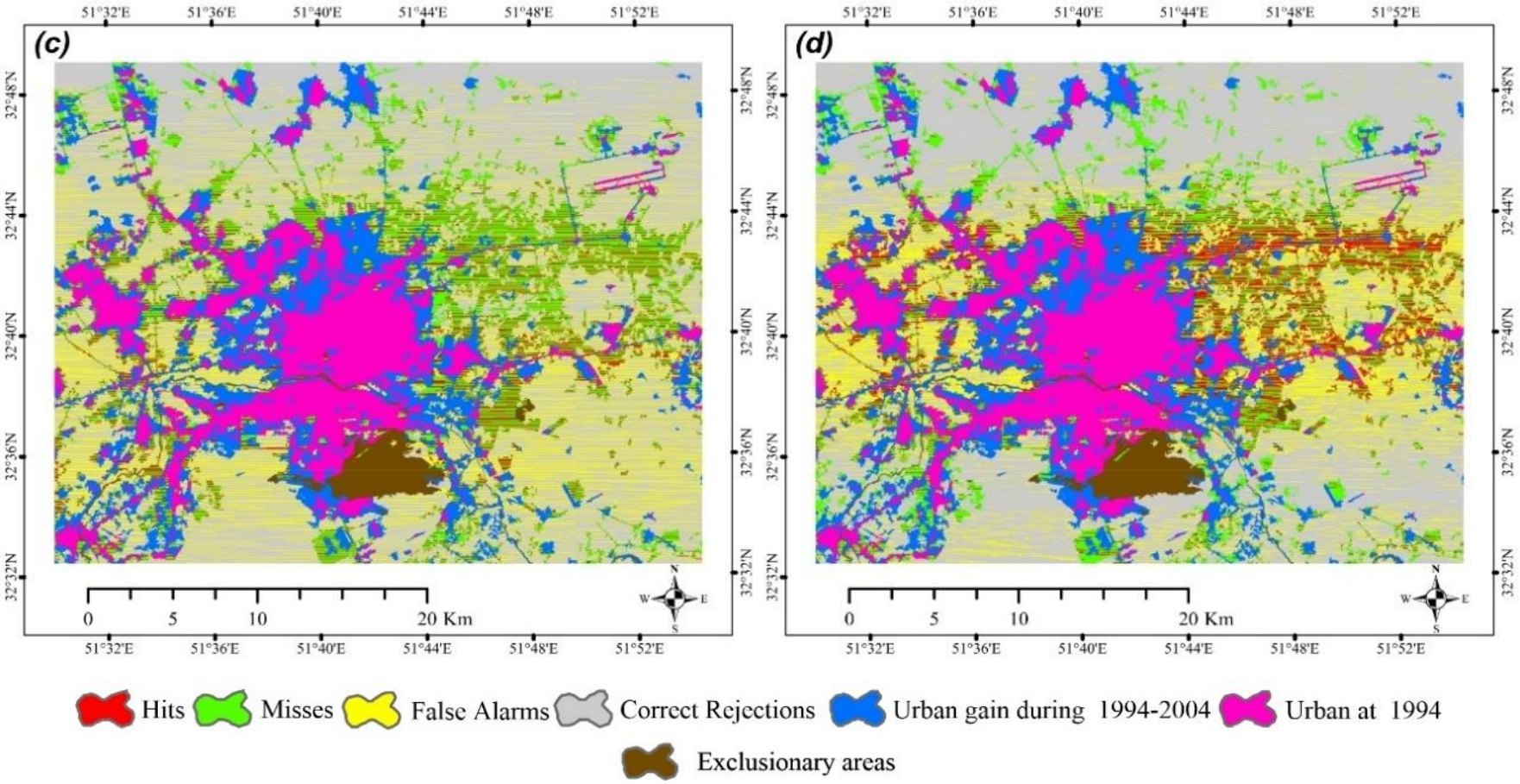
KEYWORDS
land-use change modeling; imbalance datasets; undersampling; random forest; artificial neural network; support vector machine

مدلسازی با استفاده از مدلهای دو کلاسه



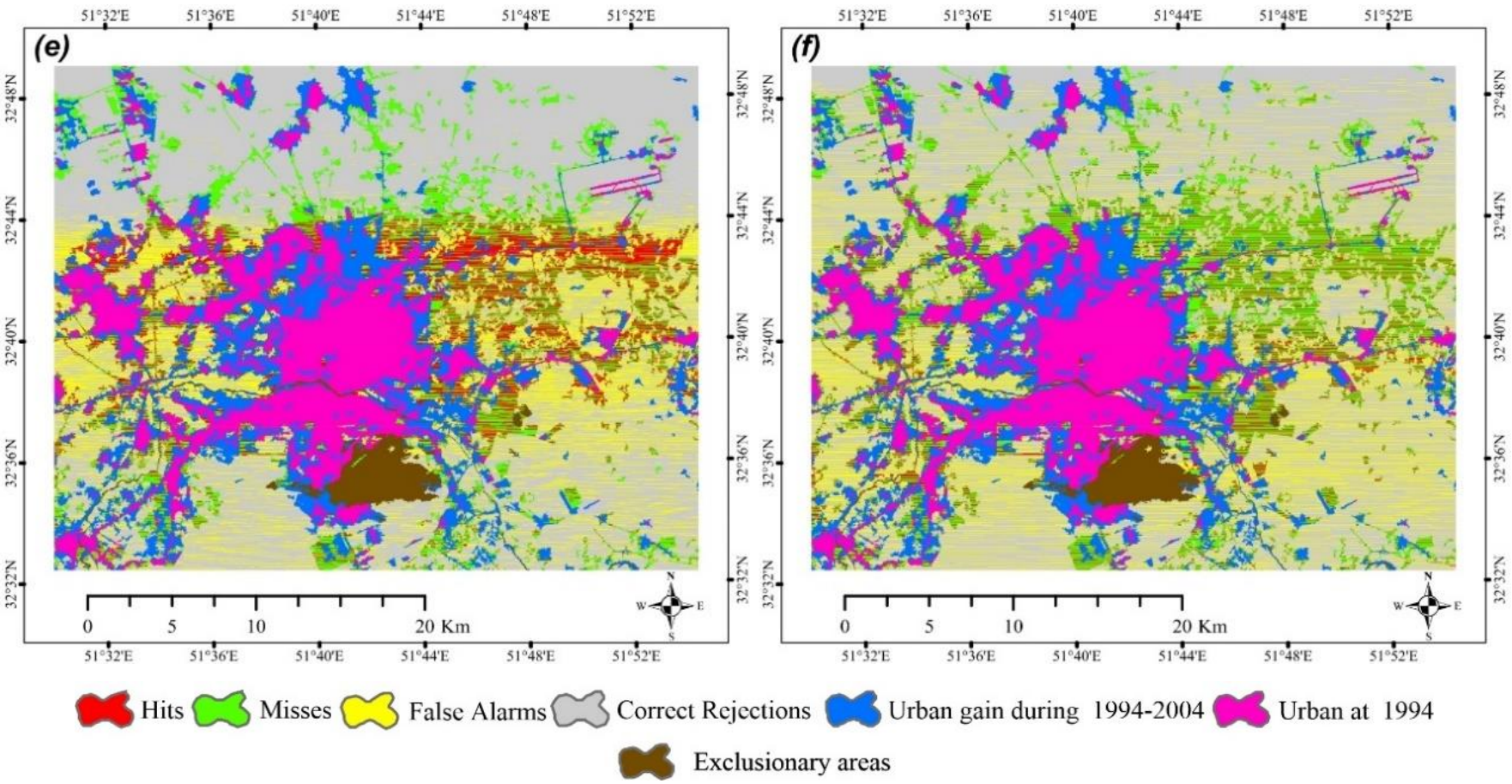
نقشه خطا محاسبه شده برای بازه زمانی دوم شهر اصفهان a) ME-ANN و b) ME-RF

مدلسازی با استفاده از مدلهای دو کلاسه



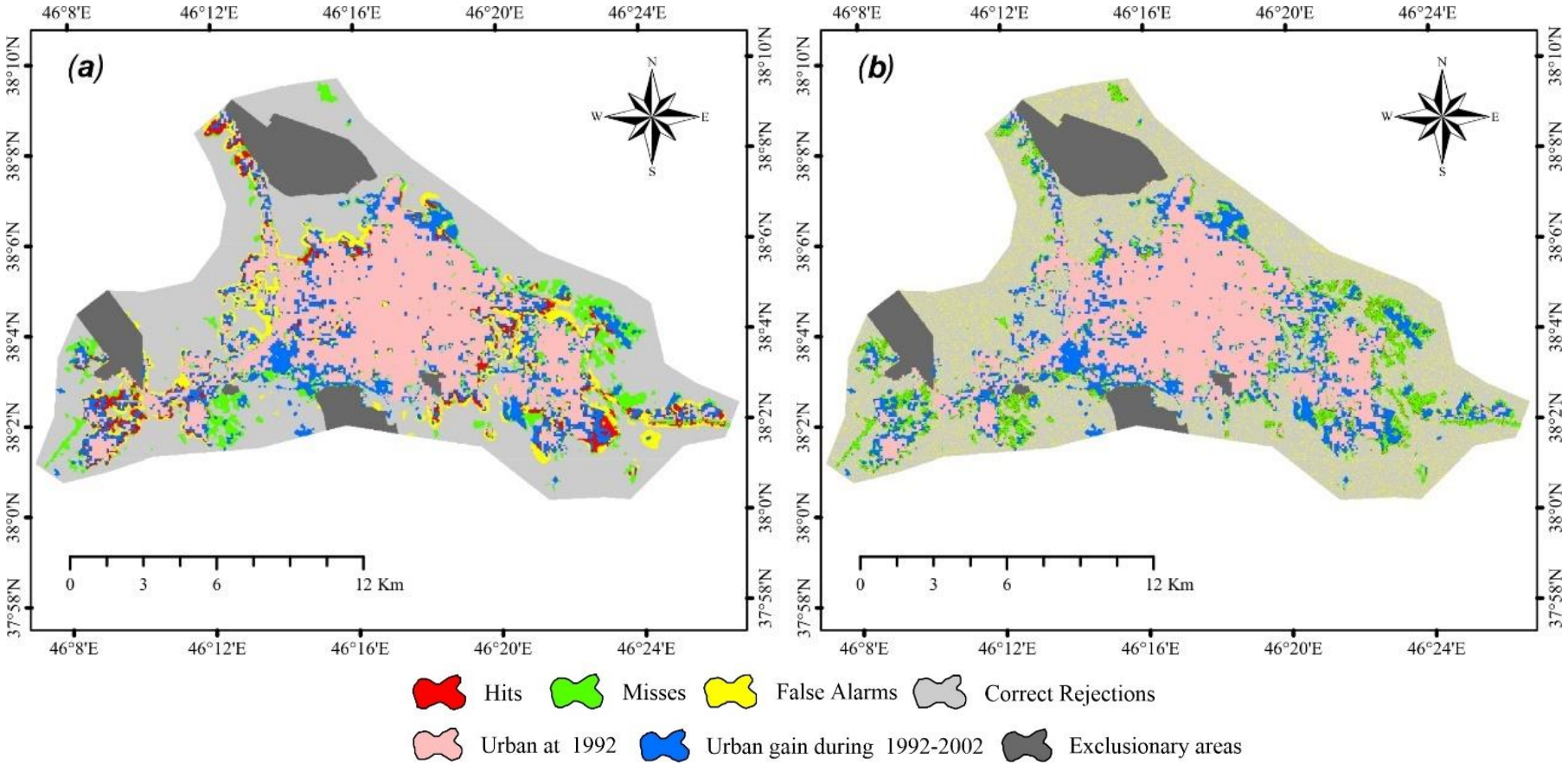
نقشه خطا محاسبه شده برای بازه زمانی دوم شهر اصفهان c) ME-SVM و d) ENFA-ANN

مدلسازی با استفاده از مدلهای دو کلاسه



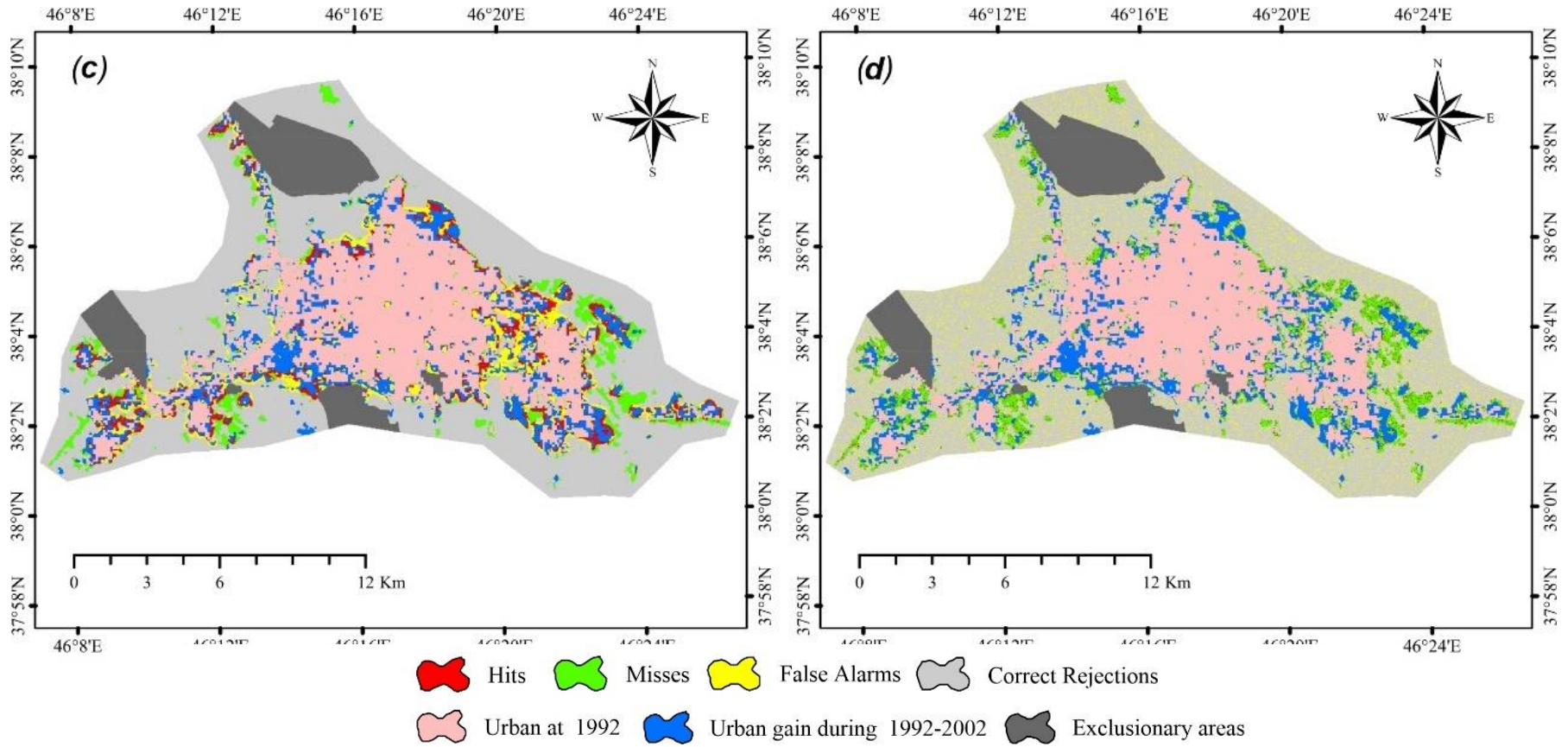
نقشه خطا محاسبه شده برای بازه زمانی دوم شهر اصفهان e) ENFA-RF و f) ENFA-SVM

مدلسازی با استفاده از مدلهای دو کلاسه



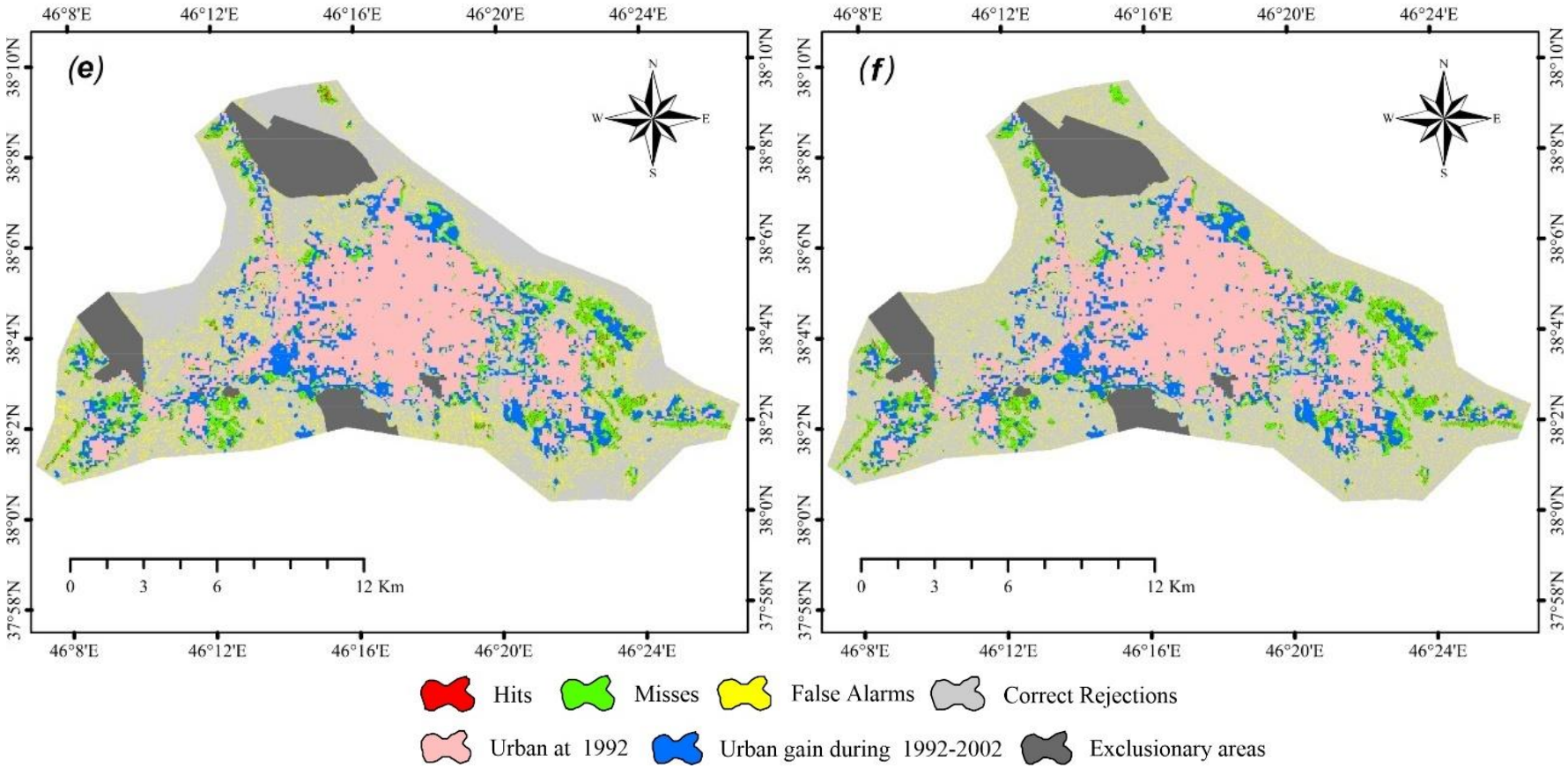
نقشه خطا محاسبه شده برای بازه زمانی دوم شهر تبریز a) ME-ANNو b) ME-RF

مدلسازی با استفاده از مدلهای دو کلاسه



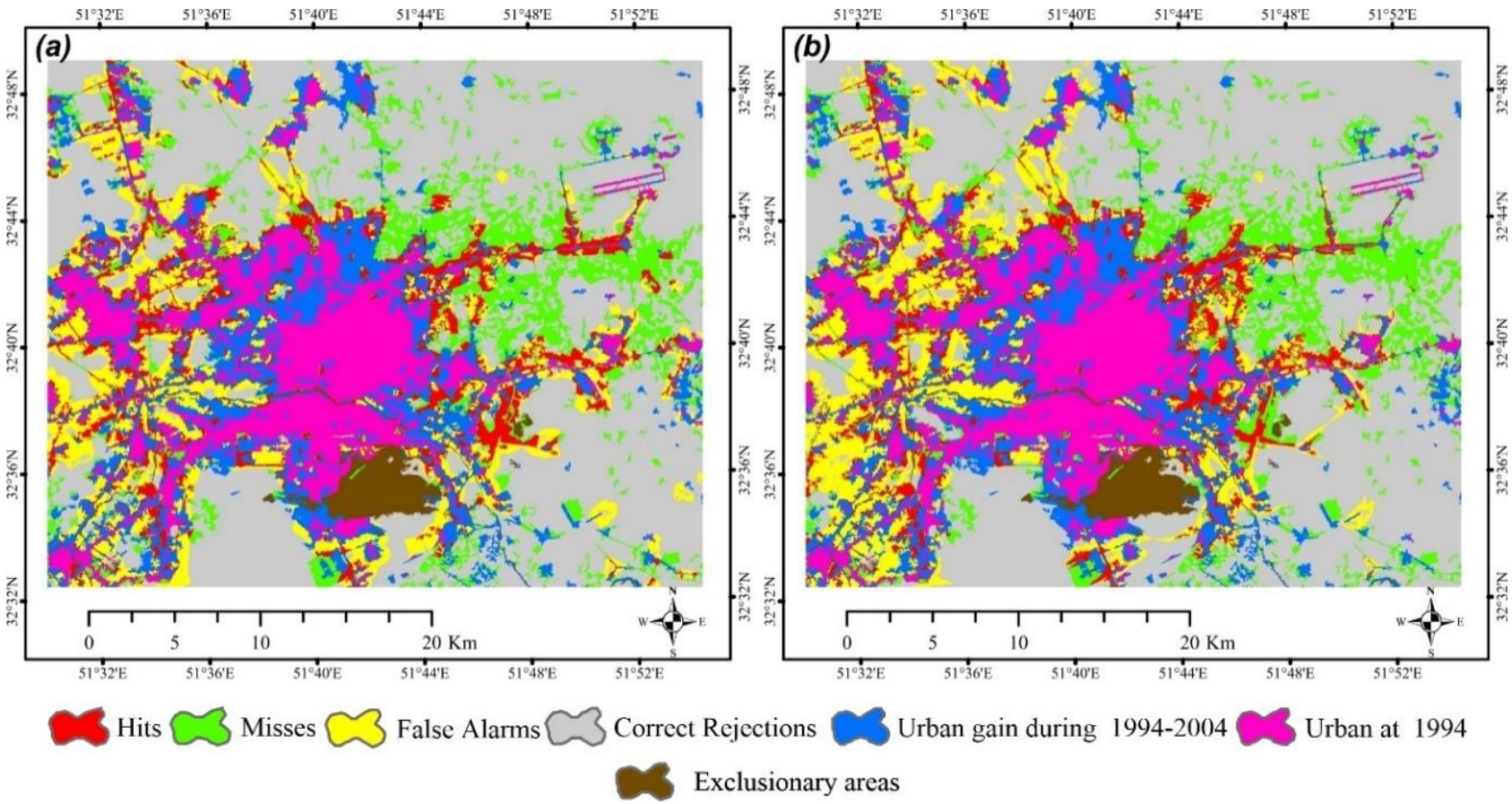
نقشه خطا محاسبه شده برای بازه زمانی دوم شهر تبریز c) ME-SVM و d) ENFA-ANN

مدلسازی با استفاده از مدلهای دو کلاسه



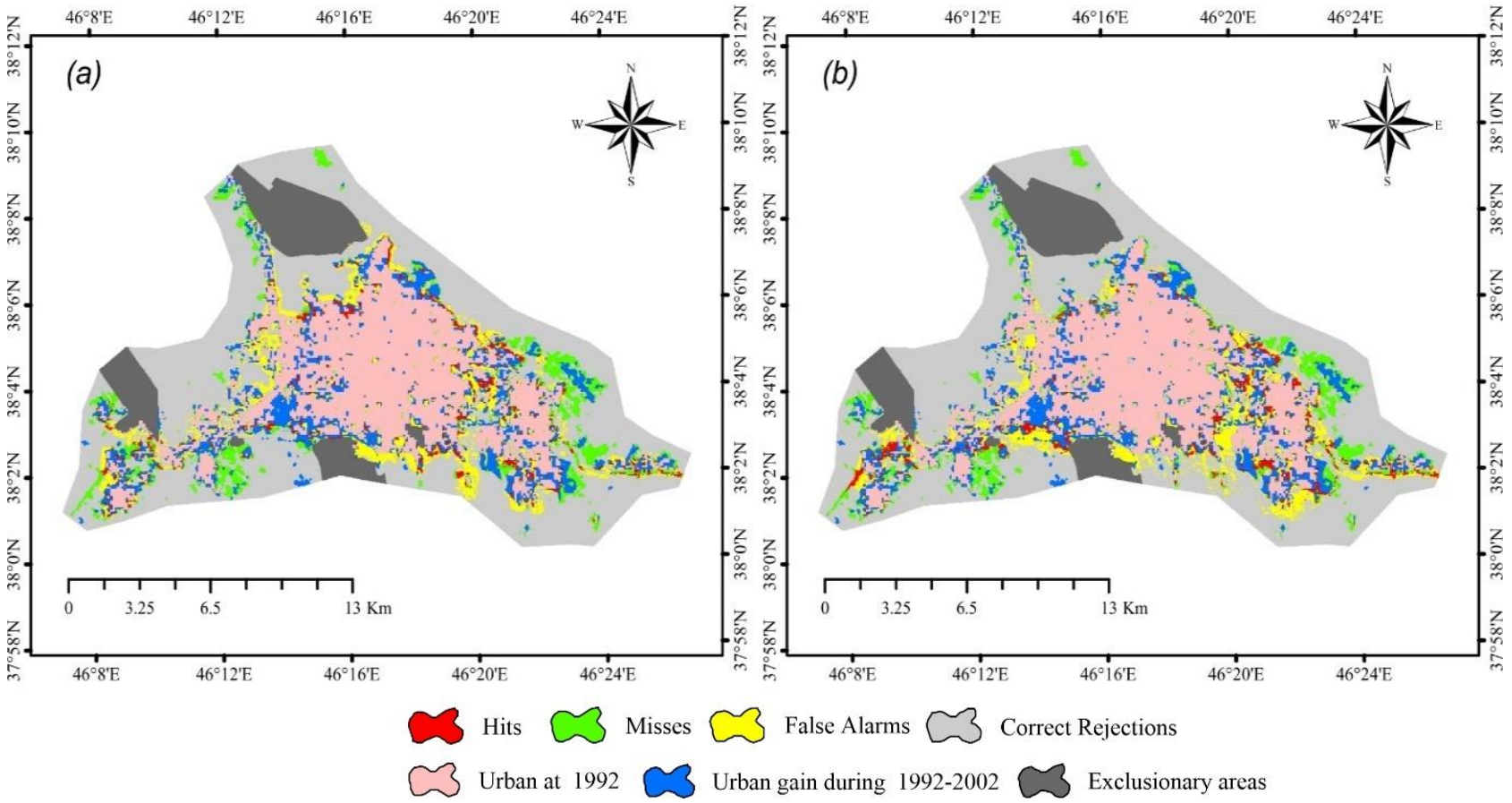
نقشه خطا محاسبه شده برای بازه زمانی دوم شهر تبریز e) ME-SVM و f) ENFA-ANN

مدلسازی تک کلاسه بر اساس آنتروپی بیشینه و تحلیل عامل

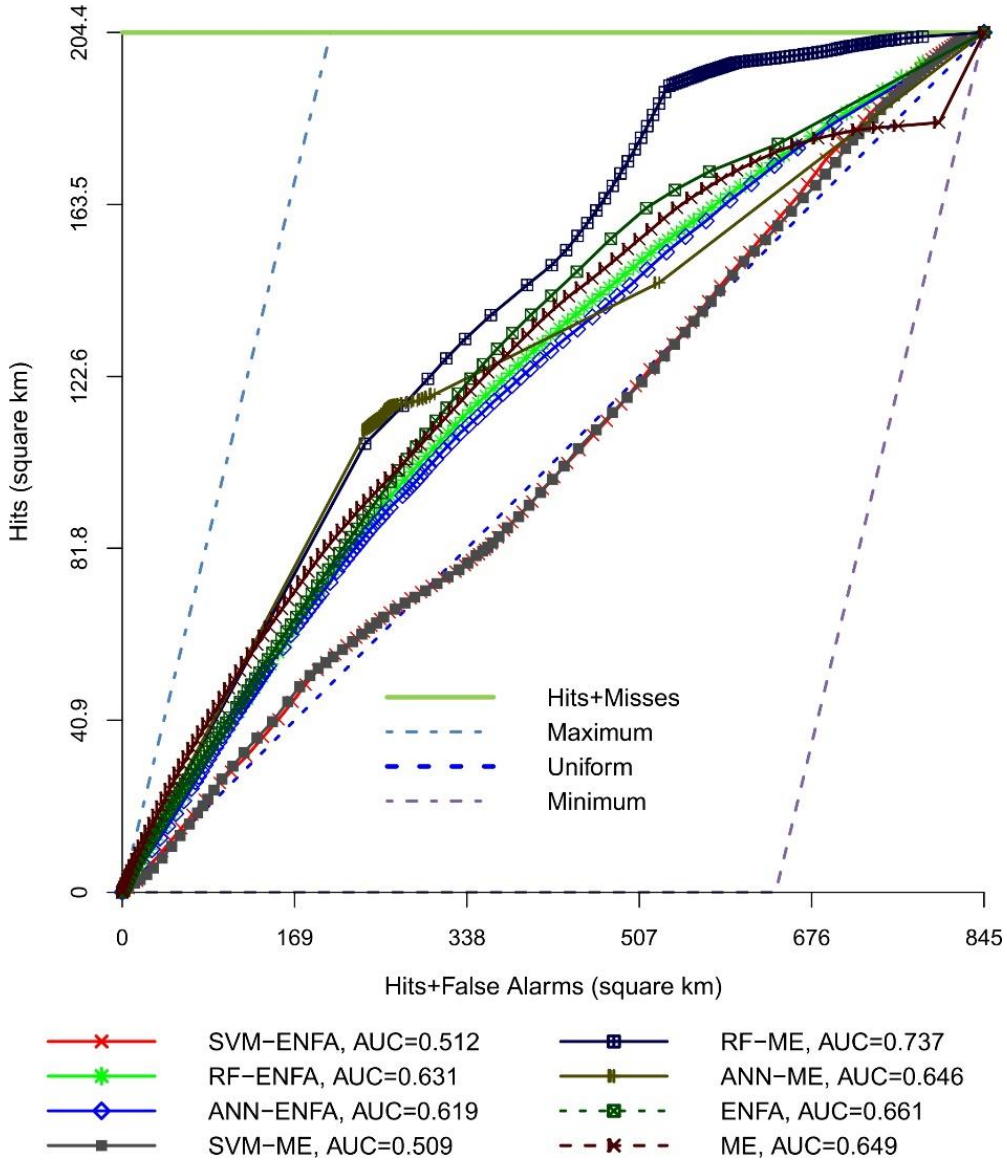


نقشه خطا محاسبه شده برای بازه زمانی دوم شهر اصفهان a) ME و b) ENFA

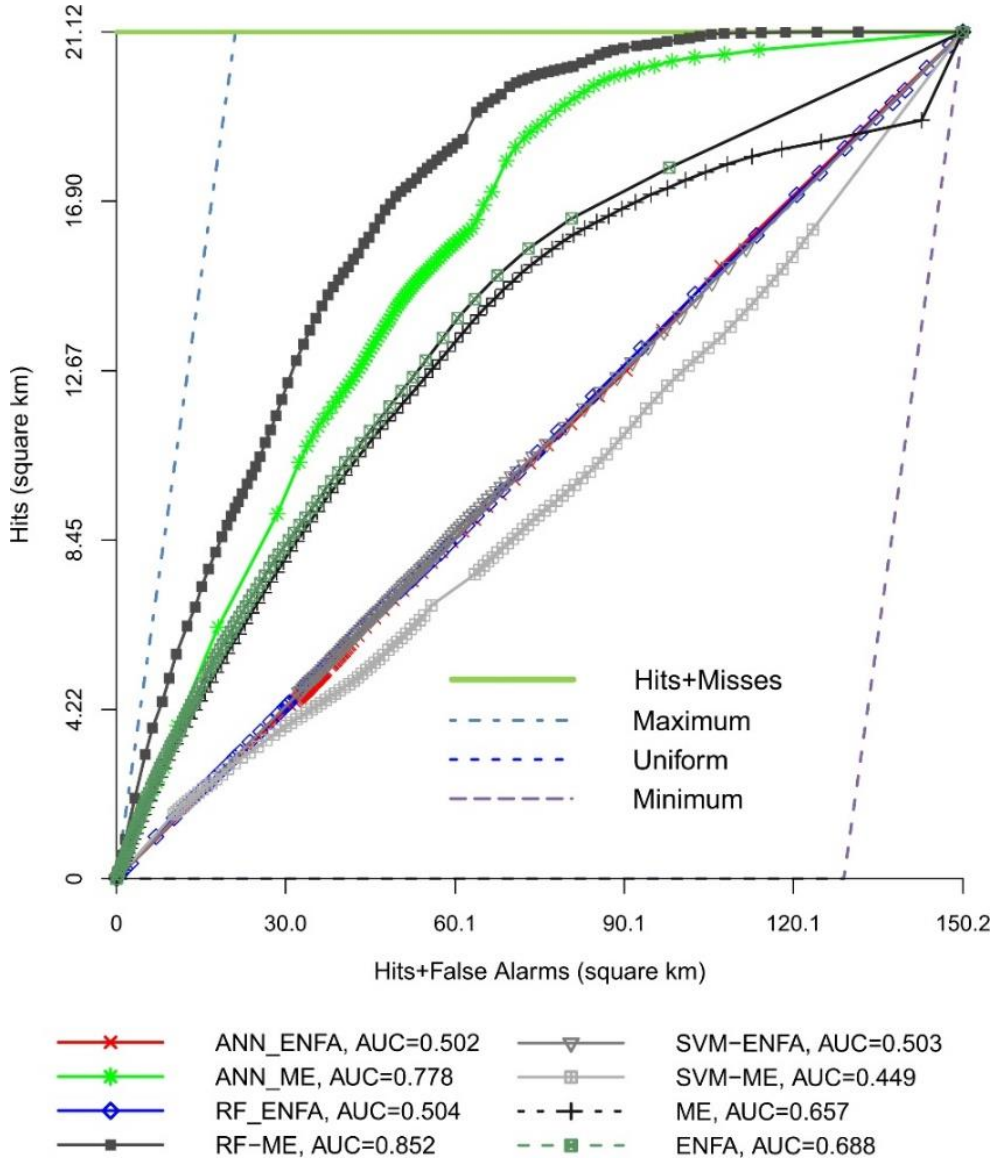
مدلسازی تک کلاسه بر اساس آنتروپی بیشینه و تحلیل عامل



نقشه خطا محاسبه شده برای بازه زمانی دوم شهر تبریز a) ME و b) ENFA



نمودار شاخص عامل کلی برای مدل های توسعه داده شده در مرحله اعتبارسنجی برای شهر اصفهان



نمودار شاخص عامل کلی برای مدل های توسعه داده شده در مرحله اعتبارسنجی برای شهر تبریز

- مناطق مورد مطالعه و داده ها
- نتایج منطق نمونه برداری
- مدلسازی دوکلاسه مبتنی بر منطق نمونه برداری
- مدلسازی تک کلاسه
- اعتبارسنجی نتایج
- نتایج مدلسازی مبتنی بر هزینه
- اعتبارسنجی نتایج

اصفهان	PA		FoM
ME	0.41		0.16
ENFA	0.38		0.18
ME-RF	0.49		0.28
ME-ANN	0.51		0.20
ME-SVM	0.27		0.07
ENFA-RF	0.37		0.08
ENFA-ANN	0.36		0.08
ENFA-SVM	0.27		0.08

GISCIENCE & REMOTE SENSING
2023, VOL. 60, NO. 1, 2222980
https://doi.org/10.1080/15481603.2023.2222980

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

The use of maximum entropy and ecological niche factor analysis to decrease uncertainties in samples for urban gain models

Mohammad Ahmadlou[a], Mohammad Karimi[a] and Nadhir Al-Ansari[b]

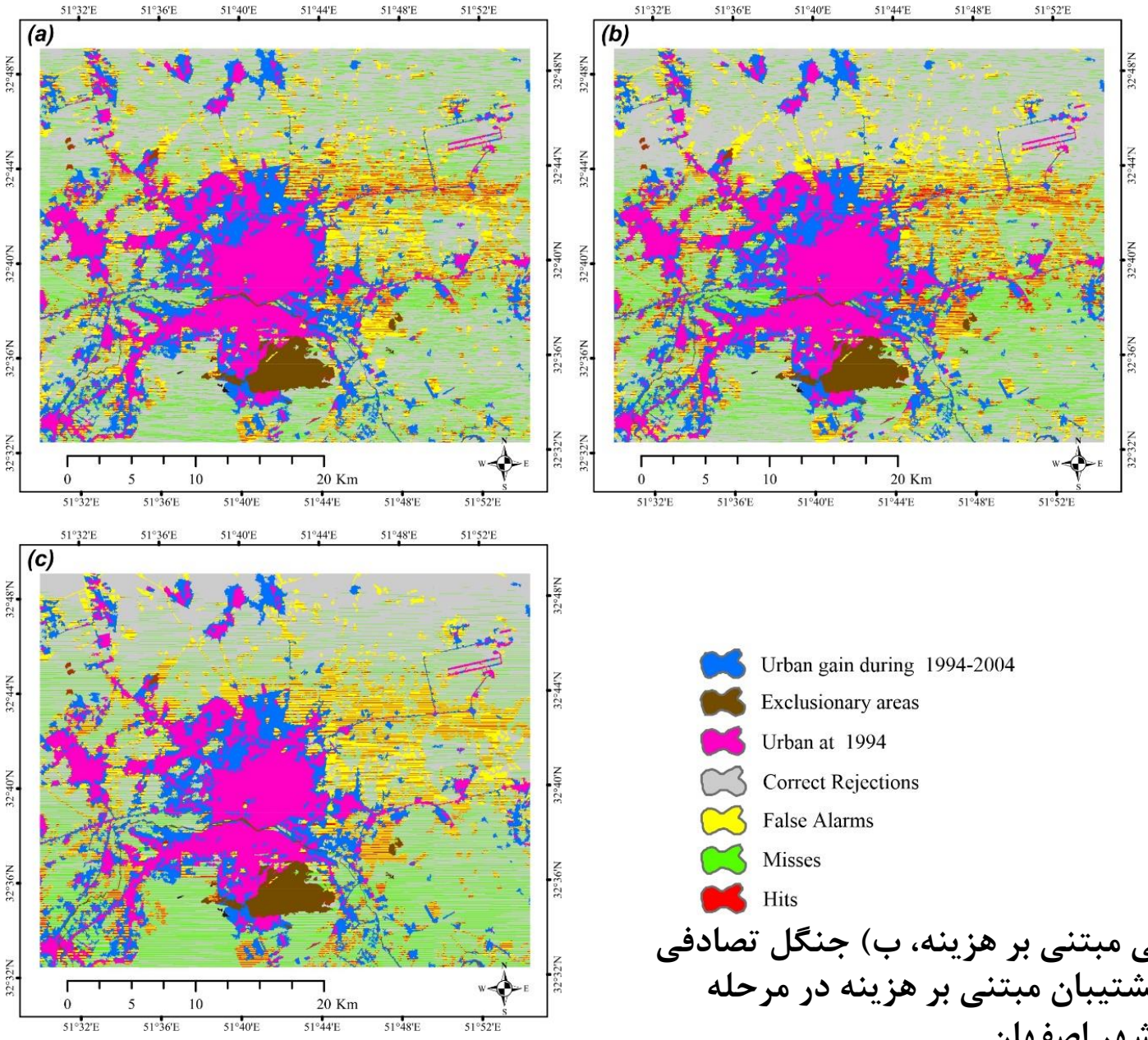
[a]GIS Department, Geodesy and Geomatics Faculty, K.N. Toosi University of Technology, Tehran, Iran; [b]Department of Civil, Environmental and Natural Resources Engineering, Lulea University of Technology, Lulea, Sweden

ABSTRACT

Uncertainty is a common problem in spatial modeling and geographical information systems (GIS). Furthermore, urban gain modeling (UGM) contains various dimensions and components of uncertainties. Data sampling is important in UGM, and may cause the results of the models to contain many uncertainties as well as affects their precision and accuracy. A poorly sampled or biased dataset can lead to inaccurate predictions and decreased performance of the models. This paper aims to present and develop novel strategies for sampling and building training datasets that can enhance the performance of data-driven models. In other words, the present study used maximum entropy (ME) and ecological niche factor analysis (ENFA) models to select pure non-change samples with minimal uncertainty for training datasets in UGM of Isfahan and Tabriz cities in Iran. The urban gain of two time intervals of 1992–2002 and 2002–2012 were used for Tabriz City and two time intervals of 1994–2004 and 2004–2014 for Isfahan City. Nine and 14 urban gain drivers were used in the UGM of Isfahan and Tabriz cities, respectively. After the ME and ENFA models produced a training dataset with change and non-change samples with the lowest uncertainty, three well-known models, namely random forest (RF), artificial neural network (ANN), and support vector machine (SVM) were used for the modeling. Moreover, the ME and ENFA models that were used to investigate the uncertainty of the sampling procedure were used as the one-class prediction models. Compared to extant studies, the proposed ME – based sampling strategy increased the area under the receiver operating characteristic curve (AUROC), figure of merit, producer's accuracy, and overall accuracy by 5.5%, 5%, 5%, and 3%, respectively, in the validation phase of Isfahan City and by 5%, 6%, 14%, and 17%, respectively, for Tabriz City. For Isfahan, the accuracies of ME (AUROC = 0.649) and ENFA (AUROC = 0.661) one – class models were closer to that of the ANN – ME (AUROC = 0.646), ANN – ENFA (AUROC = 0.619), and RF – ENFA (AUROC = 0.631) models but differed significantly from that of the RF – ME (AUROC = 0.737) model. For Tabriz, the accuracies of ME (AUROC = 0.657) and ENFA (AUROC = 0.688) one – class models were lower than that of the two class RF-ME (AUROC = 0.852), and ANN-ME (AUROC = 0.778) models. The results showed that the ME model was able to identify relatively pure non-change samples and properly remove impure non-change samples from the training dataset. This study discovered that binary models are preferable to one-class models, and showed that an optimal sampling strategy is an essential step in UGM as it can decrease uncertainty. As such, modelers must adopt efficient sampling methods.

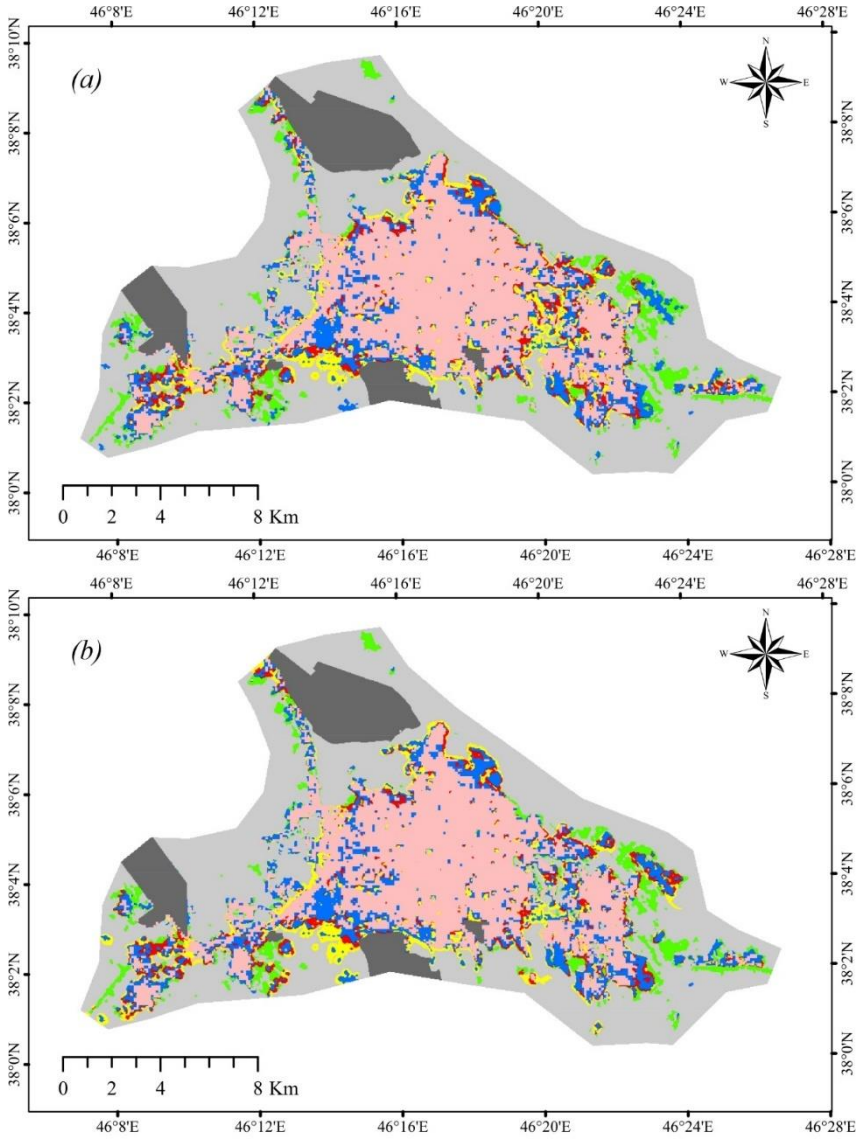
ARTICLE HISTORY
Received 27 January 2023
Accepted 05 June 2023

KEYWORDS
Uncertainty; Urban gain modeling; Data-driven models; Maximum entropy; Imbalance learning; Data mining



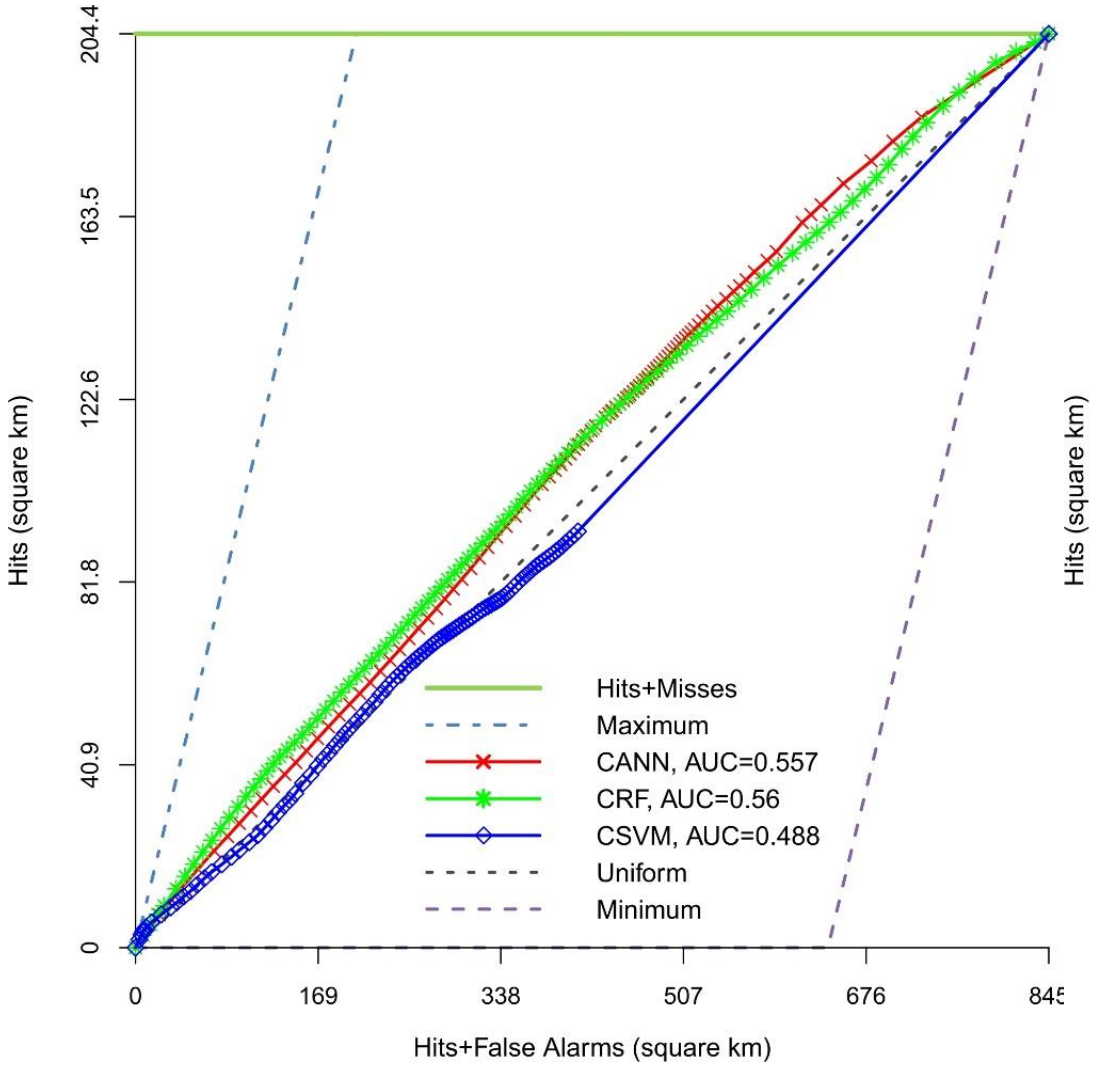
نقشه خطای سه مدل الف) شبکه عصبی مبتنی بر هزینه، ب) جنگل تصادفی مبتنی بر هزینه و ج) ماشین بردار پشتیبان مبتنی بر هزینه در مرحله اعتبارسنجی شهر اصفهان

- مناطق مورد مطالعه و داده ها
- نتایج منطق نمونه برداری
- مدلسازی دوکلاسه مبتنی بر منطق نمونه برداری
- مدلسازی تک کلاسه
- اعتبارسنجی نتایج
- نتایج مدلسازی مبتنی بر هزینه
- اعتبارسنجی نتایج

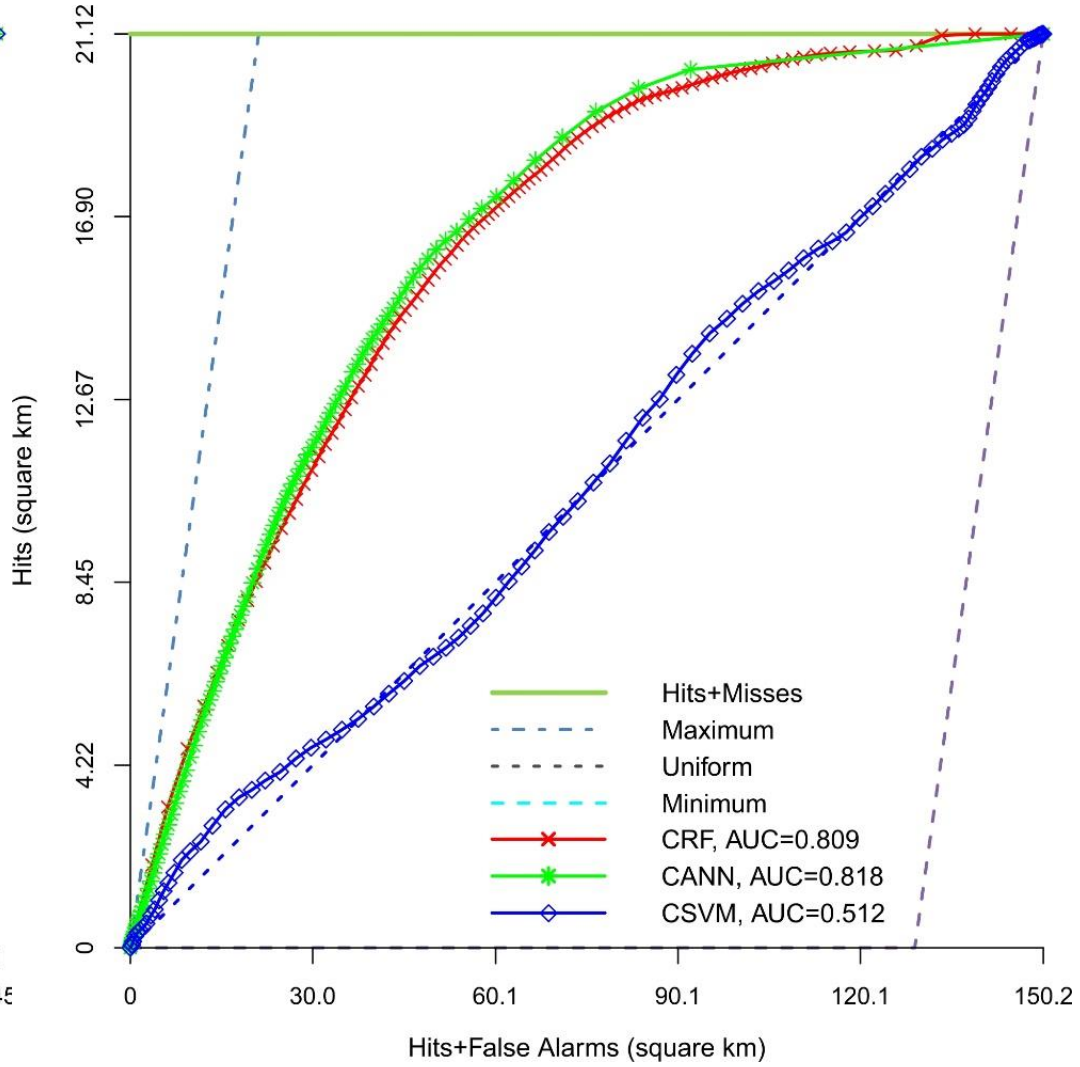


نقشه خطای سه مدل الف) شبکه عصبی مبتنی بر هزینه، ب) جنگل تصادفی مبتنی بر هزینه و ج) ماشین بردار پشتیبان مبتنی بر هزینه در مرحله اعتبارسنجی شهر تبریز

- مناطق مورد مطالعه و داده ها
- نتایج منطق نمونه برداری
- مدلسازی دوکلاسه مبتنی بر منطق نمونه برداری
- مدلسازی تک کلاسه
- اعتبارسنجی نتایج
- نتایج مدلسازی مبتنی بر هزینه
- اعتبارسنجی نتایج



نمودار شاخص عامل کلی برای مدل های حساس به هزینه در مرحله اعتبارسنجی برای شهر اصفهان



نمودار شاخص عامل کلی برای مدل های حساس به هزینه در مرحله اعتبارسنجی برای شهر تبریز

- مناطق مورد مطالعه و داده ها
- نتایج منطق نمونه برداری
- مدلسازی دوکلاسه مبتنی بر منطق نمونه برداری
- مدلسازی تک کلاسه
- اعتبارسنجی نتایج
- نتایج مدلسازی مبتنی بر هزینه
- اعتبارسنجی نتایج

GEOCARTO INTERNATIONAL
2024, VOL. 39, NO. 1, 2353252
https://doi.org/10.1080/10106049.2024.2353252

Three novel cost-sensitive machine learning models for urban growth modelling

Mohammad Ahmadlou[a], Mohammad Karimi[a], Saad Sh. Sammen[b] and Karam Alsafadi[c*]

[a]GIS Department, Geodesy and Geomatics Faculty, K.N. Toosi University of Technology, Tehran, Iran; [b]Department of Civil Engineering, College of Engineering, University of Diyala, Baqubah, Iraq; [c]Department of Geography, Faculty of Arts and Humanities, Damascus University, Damascus, Syria

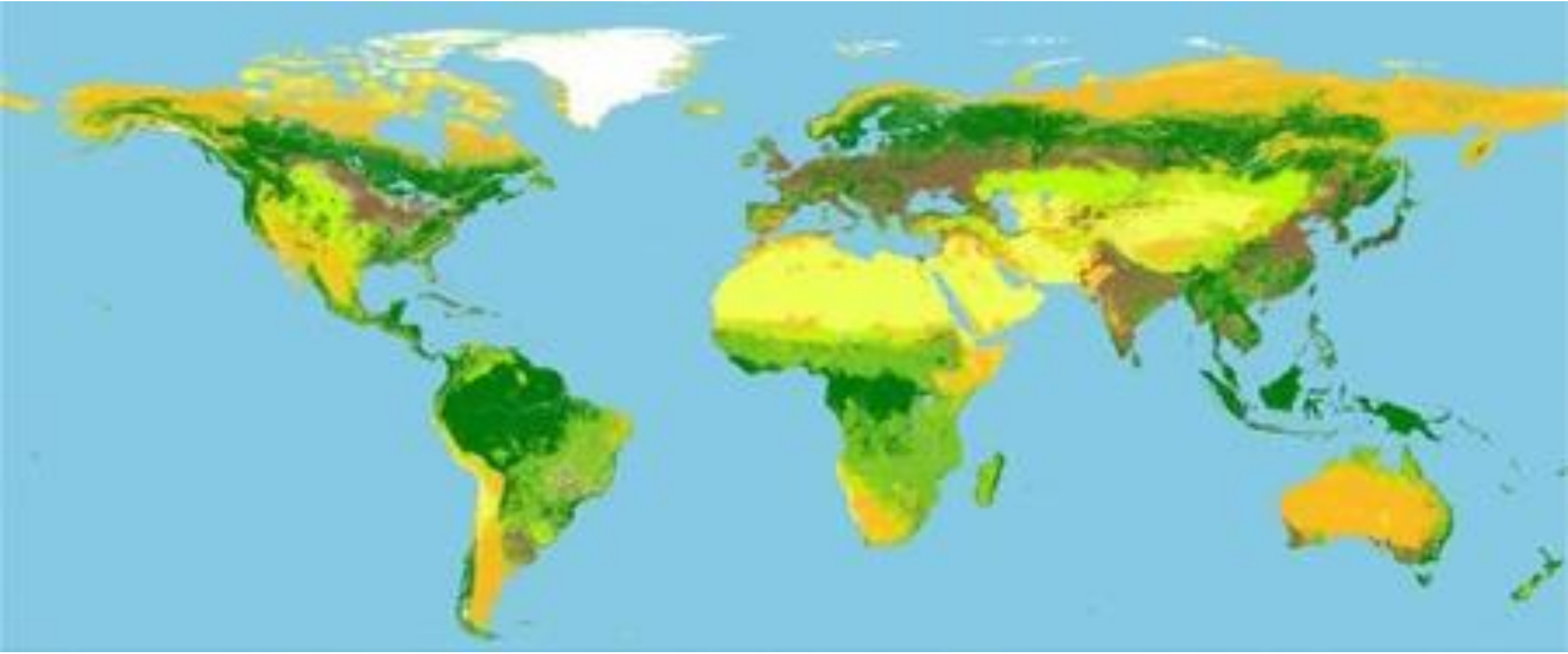
ABSTRACT

This article addresses the class imbalance problem in urban gain modelling (UGM) of Tabriz and Isfahan megacities in Iran by proposing novel cost-sensitive machine learning models, namely cost-sensitive support vector machine (CSVM), random forest (CRF) and artificial neural network (CANN). Random sampling, a frequently utilized method, fails to effectively tackle this issue by biasing models towards no change samples, which outnumber change samples. The results showed that CRF exhibited the highest accuracy (AUC = 0.560), followed by CANN (AUC = 0.557) and CSVM (AUC = 0.448) in Isfahan. In Tabriz, CRF (AUC = 0.809) and CANN (AUC = 0.818) excelled, outperforming balanced sampling models constructed with ANN, RF and SVM with the AUROC of ANN and RF boosted by 15% and 2% in validation. By emphasizing the significance of addressing class imbalance appropriately, this research highlights the improvement in modelling outcomes achievable through the cost-sensitive models especially in Tabriz case.

ARTICLE HISTORY
Received 18 September 2023
Accepted 3 May 2024

KEYWORDS
Imbalance problem; cost-sensitive algorithms; urban gain modelling; random sampling; data mining; machine learning model

اصفهان | oM
CANN | .27
CRF | .26
CSVM | .10



نتیجه گیری و پیشنهادات

- ارائه یک منطق نمونه برداری متوازن مبتنی بر آنتروپی بیشینه و تحلیل عاملی آشیان بوم شناختی برای مدلسازی رشد شهری
- استفاده از سه مدل یادگیری ماشین و داده کاوی شامل شبکه عصبی مصنوعی، ماشین بردار پشتیبان و جنگل تصادفی با منطق نمونه برداری ارائه شده
- ساخت شش مدل ترکیبی شامل ME-ANN، ME-SVM، ME-RF، ENFA-ANN، ENFA-SVM و ENFA-RF برای مدلسازی رشد شهری تبریز و اصفهان
- ارائه دو مدل تک کلاسه جدید بر اساس آنتروپی بیشینه و تحلیل عاملی آشیان بوم شناختی برای مدلسازی و مقایسه عملکرد آنها با مدلهای ساخته شده بر اساس منطق نمونه برداری پیشنهادی
- ارائه سه الگوریتم مبتنی بر هزینه شامل شبکه عصبی حساس به هزینه، ماشین بردار پشتیبان حساس به هزینه و جنگل تصادفی حساس به هزینه برای مقابله با مسئله نامتوازنی
- مقایسه دو رویکرد سطح داده و سطح الگوریتم برای مقابله با مسئله نامتوازنی داده رشد شهری

- افزایش قابل توجه دقت مدلسازی با منطق نمونه برداری توسعه داده شده مبتنی بر آنتروپی بیشینه در مدلهای جنگل تصادفی و شبکه عصبی مصنوعی و کاهش اندازه مجموعه داده‌های آموزشی و بارمحاسباتی
- افزایش دقت مدلسازی با رویکرد نمونه برداری توسعه داده شده بر اساس مدل تحلیل عاملی آشیان بوم شناختی در برخی از مدل ها
- عملکرد بهتر مدل های دو کلاسه مبتنی بر منطق نمونه برداری توسعه داده شده نسبت به مدل های تک کلاسه
- عملکرد ضعیف مدلهای مبتنی بر هزینه ارائه شده در مدلسازی رشد شهری اصفهان
- افزایش قابل توجه دقت مدلسازی با دو الگوریتم مبتنی بر هزینه شبکه عصبی مصنوعی و جنگل تصادفی در مدلسازی رشد شهری تبریز

نتیجه گیری

پیشنهادات

- عملکرد بسیار ضعیف مدل ماشین بردار پشتیبان توسعه داده شده در هر دو رویکرد
- عملکرد بهتر مدل جنگل تصادفی ساخته شده با منطق نمونه برداری مبتنی بر آنتروپی بیشینه نسبت به مدل جنگل تصادفی مبتنی بر هزینه در شهر تبریز
- عملکرد بهتر مدل شبکه عصبی مبتنی بر هزینه نسبت به مدل شبکه عصبی ساخته شده با منطق نمونه برداری مبتنی بر آنتروپی بیشینه در شهر تبریز


نتیجه گیری
پیشنهادات
مسئله نامتوازنی و ناخالصی نمونه‌های عدم تغییر در مدل‌سازی تغییرات کاربری اراضی چندگانه بسیار پیچیده است، زیرا غیر از مسئله نامتوازنی بین کلاس‌های تغییر و عدم-تغییر، این مسئله و چالش بین کلاس‌های تغییر کاربری مختلف نیز رخ می‌دهد. بنابراین، محققان در مطالعات آتی می‌توانند منطق نمونه‌برداری پیشنهادی در این پژوهش را به مدلسازی تغییرات کاربری چندگانه تعمیم دهند.
در مطالعه حاضر دو رویکرد نمونه‌برداری مبتنی بر نمونه‌برداری کمینه استفاده شد که اقدام به متوازن-سازی مجموعه داده آموزشی از طریق کم کردن تعداد نمونه‌های عدم تغییر کرد. در مطالعات آتی محققان می‌توانند از رویکردهای مبتنی بر نمونه-برداری بیشینه که اقدام به تولید نمونه‌های مصنوعی از کلاس اقلیت به منظور متوازن‌سازی مجموعه داده آموزشی کنند.
در تحقیقات آتی سعی خواهد شد یک رویکرد ترکیبی از دو رویکرد سطح داده و سطح الگوریتم برای مقابله با مسئله نامتوازنی ارائه شود. به نحوی که ابتدا با رویکردهای سطح داده تا حد امکان مسئله نامتوازنی تقلیل یافته سپس برای مقابله با نامتوازنی باقی-مانده از الگوریتم‌های حساس به هزینه استفاده شود.
استفاده از رویکردهای سطح الگوریتم برای مقابله با مسئله نامتوازنی از مدلی به مدل دیگر متفاوت و وابسته به ساختار الگوریتم است. لذا پیشنهاد می‌شود در مطالعات آتی اقدام به وزندار کردن سایر مدل‌ها به منظور مدلسازی داده-های رشد شهری یا تغییرات کاربری اراضی با نرخ تغییر مختلف شود.
منطق نمونه‌برداری پیشنهادی قابلیت استفاده در سایر حیطه مربوط به مدلسازی مکانی را نیز دارا می‌باشد. لذا پیشنهاد می‌شود این رویکرد در سایر حیطه‌های تحقیقاتی نیز استفاده شده و کارایی آنها سنجیده شود.

مقالات چاپ شده

- ✓ Ahmadlou, Mohammad, Mohammad Karimi, and Nadhir Al-Ansari. "The use of maximum entropy and ecological niche factor analysis to decrease uncertainties in samples for urban gain models." *GIScience & Remote Sensing* 60.1 (2023): 2222980.

- ✓ Ahmadlou, Mohammad, Mohammad Karimi, and Robert Gilmore Pontius Jr. "A new framework to deal with the class imbalance problem in urban gain modeling based on clustering and ensemble models." *Geocarto International* 37.19 (2022): 5669-5692.

- ✓ Ahmadlou, Mohammad, et al. "Three novel cost-sensitive machine learning models for urban growth modelling." Geocarto International 39.1 (2024): 2353252.

ممنون از توجه شما

مرداد ۱۴۰۳