نشریه علمی - پژوهشی علوم و فنون نقشه برداری، دوره پنجم ، شماره ۳ ، بهمن ماه ۱۳۹۴

# ارزیابی دقت توصیفی عوارض در اطلاعات مکانی مردم‌گستر

**بهزاد واحدی طرقبه*[۱]، علی اصغر آل شیخ[۲]**

[۱] کارشناس ارشد سیستم‌های اطلاعات مکانی- دانشکده مهندسی نقشه‌برداری - دانشگاه صنعتی خواجه نصیر الدین طوسی
behzad@geog.ucsb.edu

[۲] استاد گروه سیستم‌های اطلاعات مکانی- دانشکده مهندسی نقشه‌برداری- دانشگاه صنعتی خواجه نصیر الدین طوسی (عضو قطب علمی مهندسی فناوری اطلاعات مکانی)
alesheikh@kntu.ac.ir

(تاریخ دریافت تیر۱۳۹۴، تاریخ تصویب آذر ۱۳۹۴)

## چکیده

از زمان پیدایش مفهوم اطلاعات مکانی مردم‌گستر (داوطلبانه)[۱] کیفیت این اطلاعات به عنوان بزرگترین مشکل آن معرفی شده است. بنابراین تا کنون تحقیقات مختلفی به بررسی کیفیت داده‌های مردم‌گستر پرداخته و سعی در برآورد کیفیت این اطلاعات داشته اند. اما در این تحقیقات به دقت توصیفی کمتر از سایر المان‌های کیفیت پرداخته شده است؛ در حالیکه این المان در آنالیزهای گوناگون مکانی و کاربردهای مختلف اطلاعات مردم گستر از اهمیت بالایی برخوردار است. بنابراین در این تحقیق با استفاده از یک روش جدید و با استفاده از الگوریتم Levenshtein به همراه پیش پردازش های متنی، دقت توصیفی عوارض مردم‌گستر (در قالب نام عارضه) با مقایسه آنها با عوارض مرجع مورد بررسی قرار می‌گیرد. برای محاسبه دقت توصیفی فرض می‌شود که بین عوارض مرجع و مردم‌گستر تناظریابی انجام شده است. منطقه مورد مطالعه این تحقیق شهر تهران است و از داده‌های تولیدی شهرداری تهران به عنوان مجموعه داده مرجع و از داده‌های سایت OpenStreetMap به عنوان مجموعه داده مردم‌گستر استفاده شده است. طبق نتایج حاصل، ۳۳ درصد از عوارض مردم‌گستر دارای نام، نام صحیح، ۴۴ درصد از آن‌ها نام تقریباً صحیح و ۲۳ درصد باقیمانده نام نادرست دارند و دقت توصیفی کل داده‌های مردم‌گستر برابر ۷۷ درصد می‌باشد.

**واژگان کلیدی:** اطلاعات مکانی مردم‌گستر، دقت توصیفی، الگوریتم Levenshtein، کیفیت اطلاعات مکانی، تناظریابی، OpenStreetMap

* نویسنده رابط

۱ Volunteered Geographic information

## ۱- مقدمه

پیشرفت تکنولوژی در دهه اول هزاره سوم به خصوص در زمینه‌های وب (Web 2.0) و دستگاه‌های تعیین موقعیت همراه و همزمان با آن، رشد روزافزون نیاز مصرف کنندگان عادی به اطلاعات مکانی باعث به وجود آمدن نوع جدیدی از اطلاعات مکانی شد که در آن مصرف کنندگان، خود به تولید اطلاعات مکانی پرداخته و از مصرف کننده صِرف به تولیدکننده-مصرف کننده[۱] تبدیل شدند [۱]. Goodchild در سال ۲۰۰۷ این نوع جدید اطلاعات مکانی را اطلاعات مکانی مردم‌گستر نامید [۲]. وی عنوان کرد که هر یک از انسان‌ها می‌توانند به عنوان یک حسگر عمل کنند. در نظر او دنیا متشکل از ۷ میلیارد حسگر متحرک است که می‌توانند داده‌های مکانی منحصر به فردی از محیط پیرامون خود تولید کنند. تعریف رسمی اطلاعات مکانی مردم‌گستر بدین صورت می‌باشد: " *بهره‌برداری از ابزار برای ایجاد، جمع‌آوری و انتشار داده‌های جغرافیایی که به طور داوطلبانه توسط افراد تولید شده است*" [۲].

از زمان ارائه این تعریف تا کنون، تحقیقات مختلفی درباره اطلاعات مکانی مردم‌گستر به انجام رسیده و در بسیاری از آنها کیفیت این اطلاعات به عنوان بزرگ‌ترین مشکل آن معرفی شده است [۳]. از آنجایی که تولید کنندگان این اطلاعات عموما مردم عادی بوده و هیچ تخصصی در زمینه اطلاعات مکانی، جغرافیا، و یا سایر علوم مرتبط با آن ندارند، چنین مشکلی طبیعی به نظر می‌رسد. ضمنا عدم وجود مکانیزم‌های کنترل کیفیت و یا کارایی پایین آنها در صورت وجود، در اغلب پروژه‌های مردم‌گستر این مشکل را تشدید می‌کند [۴]. بنابراین تا کنون تحقیقات مختلفی به بررسی کیفیت داده‌های مردم‌گستر پرداخته و سعی در برآورد کیفیت این اطلاعات داشته اند.

این تحقیقات را در دو گروه می‌توان دسته بندی کرد: (۱) بررسی کیفیت اطلاعات مردم‌گستر از طریق مقایسه آنها با اطلاعات مرجع (۲) بررسی ماهیت خود داده‌های مردم‌گستر و ارزیابی پارامترهای کیفیت آن‌ها [۵]. در تحقیقات دسته اول (که پژوهش حاضر هم در این دسته قرار می‌گیرد) معمولا چند المان برای کیفیت در نظر گرفته می‌شود و با مقایسه داده‌های مرجع و مردم‌گستر، یک برآورد کمی از این المان‌ها محاسبه می‌گردد. مطرح‌ترین المان هایی که برای این منظور در نظر گرفته می‌شوند عبارتند از: تمامیت، سازگاری منطقی، دقت مکانی، دقت زمانی و دقت توصیفی [۶]. اما علی الرغم اهمیت فراوان، به دقت توصیفی کمتر از سایر المان‌ها پرداخته شده است [۷]. از آنجاییکه ماهیت اغلب اطلاعات توصیفی به صورت غیر عددی است، از این‌رو ارزیابی کیفیت آن دشوارتر از سایر المان‌های کیفیت است.

علاوه بر این، بر اساس تحقیقات نگارندگان، در پژوهش‌هایی که به بررسی کیفیت اطلاعات مکانی مردم‌گستر در کشور ایران پرداخته اند به دقت توصیفی توجهی نشده یا کمتر توجه شده است [۱۵و۱۶]. در حالی که دقت اطلاعات توصیفی یکی از فاکتورهای مهم در استفاده بهینه از داده‌های مردم‌گستر است. برای مثال آنالیزهایی مثل مسیریابی یا حتی آنالیزهای ساده ای همچون انتخاب یا تجمیع عوارض برمبنای فیلد اطلاعات توصیفی به طور مستقیم با دقت اطلاعات توصیفی در ارتباط هستند. بنابراین هدف این تحقیق این است که با مقایسه اطلاعات مردم‌گستر با اطلاعات مرجع از طریق یک روش خودکار، دقت توصیفی عوارض مردم‌گستر را محاسبه کند. برای این منظور شهر تهران به عنوان منطقه مورد مطالعه انتخاب، و با استفاده از یک حالت بهبود یافته الگوریتم Levenshtein [۲۲] دقت و تمامیت اطلاعات توصیفی درون این منطقه محاسبه شده است.

ساختار ادامه این مقاله بدین صورت می‌باشد: در بخش دوم پیشینه‌ای از تحقیقات انجام شده در زمینه ارزیابی کیفیت اطلاعات مردم‌گستر و به ویژه ارزیابی دقت توصیفی بیان می‌شود. سپس در بخش سوم روش پیشنهاد شده برای محاسبه دقت توصیفی ارائه شده و در بخش چهارم نتایج حاصل از پیاده سازی روش پیشنهادی برای منطقه مورد مطالعه بیان می‌شوند. بخش پنجم هم به بیان جمع بندی و نتایج حاصل از تحقیق اختصاص دارد.

## ۲- پیشینه تحقیق

تا کنون تحقیقات فراوانی با مقایسه اطلاعات مردم‌گستر و مرجع به ارزیابی کیفیت اطلاعات مردم‌گستر پرداخته اند. بسیاری از این محققان از داده‌های سایت OpenStreetMap (OSM)، که در طول چند سال اخیر به یکی از موفق ترین و بزرگترین نمونه‌های پروژه‌های مردم‌گستر تبدیل شده است

[۱] Produser (from producer and user)

[۸و۹]، به عنوان منبع داده مردم گستر استفاده کرده اند. به طور مثال، Haklay داده‌های OSM را با داده‌های سازمان نقشه برداری انگلستان مقایسه کرده است. وی نقشه‌های مرجع جاده‌های انگلستان را با نقشه‌های موجود در OSM مقایسه کرده اما در این مقایسه تنها یک نوع از جاده‌ها را مورد بررسی قرار داده است. او با فرض اینکه طبقه‌بندی جاده‌ها در OSM درست انجام شده، از عوارض فاقد برچسب (اطلاعات توصیفی) و یا با برچسب غلط صرف‌نظر کرده است. چنین تصمیمی باعث کم شدن دقت کار شده است [۱۰].

Kounadi کیفیت داده‌های مردم‌گستر را در یک ناحیه به مساحت ۲۵ کیلومتر مربع در آتن ارزیابی کرده است. نتایج کار وی حاکی از این است که در منطقه مورد مطالعه تمامیت داده نسبتاً خوب، اما تمامیت اطلاعات توصیفی نسبتاً پایین است. همچنین دقت هندسی بالاست و دقت توصیفی در سطح قابل قبولی قرار دارد [۱۱].

Girres و Touya در فرانسه کیفیت داده‌های مردم‌گستر OSM را با مقایسه آن‌ها با داده‌های سازمان نقشه برداری فرانسه ارزیابی کرده اند. آن‌ها این مقایسه را برای چند منطقه که به صورت تصادفی انتخاب شده بودند انجام داده اند و المان‌های دقت توصیفی، دقت معنایی، تمامیت، سازگاری منطقی و دقت زمانی را مورد بررسی قرار داده اند. ایشان با ارزیابی دقت توصیفی به این نتیجه رسیدند که با افزایش تعداد مشارکت کنندگان، کیفیت اطلاعات توصیفی از نظر کمی بالاتر می‌رود. دقت توصیفی در این تحقیق ابتدا بر اساس وجود یا عدم وجود اطلاعات توصیفی ارزیابی شده است که بنابر نتایج آن درصد بسیار کمی از اقلام توصیفی به جز نام در مجموعه مردم‌گستر دارای مقدار هستند. سپس اختلاف بین نام عوارض مرجع و مردم‌گستر با استفاده از الگوریتم Levenshtein محاسبه شده و فاصله بین ۱ تا ۳ قابل قبول در نظر گرفته شده است [۱۲].

Ludwig و همکاران به بررسی کیفیت VGI در آلمان پرداخته و برای این منظور داده‌های OSM را با داده‌های مرجع NAVTEQ مقایسه کرده اند. آن‌ها به ارزیابی دقت هندسی، تمامیت عوارض و تمامیت توصیفی پرداخته اند. برای ارزیابی تمامیت توصیفی، آن‌ها از دو قلم توصیفی نام اصلی و نام ثانویه عوارض استفاده کرده و آن‌ها را در دو مجموعه داده دو به دو با هم مقایسه کرده اند [۱۳]. برای مقایسه دو رشته هم از الگوریتم Levenshtein استفاده کرده اند. طبق نتایج به دست آمده، این مقاله پیشنهاد کرده است که استاندارد کردن نام عوارض می‌تواند نتایج بهتری در زمینه تناظریابی بین آنها تولید کند.

Koukoletsos و همکاران کیفیت داده‌های OpenStreetMap را در انگلستان ارزیابی کرده و برای این منظور از داده‌های سازمان نقشه برداری انگلستان استفاده کرده اند. در این پژوهش، برای سادگی کار تنها عوارض خطی در نظر گرفته شده اند و پارامترهای تمامیت، دقت هندسی و دقت توصیفی داده‌ها ارزیابی شده اند. برای ارزیابی دقت توصیفی تنها نام عوارض در نظر گرفته شده اند و تعداد حروف مشترک بین دو نام (یکی از مجموعه مردم‌گستر و دیگری از مجموعه مرجع) به عنوان معیار کیفیت در نظر گرفته شده است. اما مشکل اساسی این روش این است که در آن ترتیب حروف در نظر گرفته نمی‌شود و بنابراین ممکن است دو نام (دو رشته از حروف) کاملا متفاوت، که تصادفا حروف مشترکی با هم دارند اما جای این حروف با هم فرق دارد، به عنوان دو رشته یکسان در نظر گرفته شوند [۱۴].

در کشور ایران تا کنون چند مورد تحقیق درباره کیفیت اطلاعات مکانی مردم‌گستر به انجام رسیده است؛ از جمله می توان به تحقیقات فرقانی و همکاران [۱۵] و محمدی و همکاران [۱۶] اشاره کرد. فرقانی و همکاران با مقایسه داده‌های یک منطقه از شهر تهران در سایت OpenStreetMap با داده‌های شهرداری تهران در همین منطقه، یک شاخص کیفیت جدید برای محاسبه دقت داده‌های مردم‌گستر معرفی می‌کنند. برای این منظور، ابتدا شاخص‌هایی همچون مساحت حداقل محدوده محصور کننده و جهت بیضی خطای استاندارد برای هر دو مجموعه‌داده مرجع و مردم‌گستر محاسبه می‌شوند. سپس با استفاده از منطق فازی، میزان سازگاری این شاخص‌ها در داده‌های مردم‌گستر با داده‌های مرجع متناظر محاسبه می‌شود و به عنوان شاخص جدیدی از دقت داده‌های مردم‌گستر معرفی می‌شود [۱۵].

محمدی و همکاران [۱۶] هم برای ارزیابی کیفیت مکانی داده‌های مردم‌گستر، این داده‌ها را با داده‌های مرجع سازمان نقشه برداری مقایسه می‌کنند. سپس با ارزیابی داده‌های مردم‌گستر دارای متناظر، تعدادی پارامتر برای کیفیت مکانی معرفی می‌کنند. این پارامترها در چهار دسته تقسیم‌بندی می‌شوند که عبارتند از پارامترهای ذاتی، مکانی، زمانی و کاربر. پس از آن، با استفاده از یک روش

ترکیبی بر مبنای هوش مصنوعی، رابطه میان هر یک از این پارامترها با شاخص دقت مکانی (که یک برآورد از دقت مکانی داده است) مشخص می‌گردد. در نهایت با استفاده از این روابط، دقت مکانی داده‌های مردم‌گستر بدون متناظر محاسبه شده و از نتایج حاصل برای افزایش سازگاری منطقی داده‌های مردم‌گستر استفاده می‌شود [۱۶].

همانگونه که پیداست تحقیقات انجام شده در کشور ایران تنها دقت هندسی (مکانی) را مورد ارزیابی قرار داده اند و به دقت توصیفی توجه کمتری داشته اند. بنابراین توسعه روشی برای ارزیابی دقت توصیفی عوارض مردم‌گستر در کشور ایران ضروری به نظر می‌رسد. در بخش بعدی به روش پیشنهاد شده برای این منظور پرداخته می‌شود.

## ۳- روش پیشنهادی

هدف این تحقیق محاسبه دقت توصیفی به عنوان شاخصی از کیفیت اطلاعات مکانی مردم‌گستر است. برای برآورد کیفیت، یک مجموعه‌داده یا در واقع یک نقشه‌ی مردم‌گستر با یک مجموعه‌داده یا نقشه ای که توسط ارگان‌های رسمی نقشه برداری تهیه شده است مقایسه شده و با فرض اینکه مجموعه‌داده رسمی کامل، درست و بدون خطاست، کیفیت مجموعه‌داده مردم‌گستر بر اساس میزان اختلاف با مجموعه‌داده رسمی مشخص می‌شود. اولین و مهم‌ترین مرحله در انجام این مقایسه، تناظریابی بین دو مجموعه‌داده است چراکه برای محاسبه میزان اختلاف می‌بایست ابتدا عوارض متناظر را پیدا کرد. پس از یافتن عوارض متناظر، می‌بایست برای هر عارضه در مجموعه مردم‌گستر، دقت توصیفی را محاسبه نمود. برای تناظریابی بین دو مجموعه داده مکانی روش‌های مختلفی وجود دارد که می توانند شامل روش‌های رستر مبنا و یا عارضه مبنا باشند [۱۴][۱۶][۱۸]. اما از آنجایی که تمرکز این تحقیق بر محاسبه دقت توصیفی است، به جزئیات روش تناظریابی پرداخته نمی‌شود. اما کلیات این روش به شرح زیر است:

روش تناظریابی مورد استفاده در این تحقیق یک الگوریتم پنج مرحله‌ای است. این پنج مرحله به ترتیب روی داده‌ها اعمال شده و در هر مرحله برای تعدادی از عوارض یک مجموعه‌داده (مردم‌گستر یا مرجع)، عارضه متناظر در مجموعه‌داده مقابل پیدا می‌شود. مراحل اول تا چهارم روی مجموعه مردم‌گستر و مرحله پنجم روی مجموعه مرجع اعمال می‌شود. بدین معنا که در مراحل اول تا چهارم، یک عارضه از مجموعه مرجع انتخاب شده و در مجموعه مردم‌گستر، با اعمال شروط و قیودی مشخص، عارضه متناظر برای عارضه انتخاب شده جستجو می‌شود. در مرحله پنجم عوارض مجموعه مردم‌گستر انتخاب شده و در مجموعه مرجع به دنبال عارضه متناظر جستجو می‌شود. جزییات این روش توسط واحدی ارائه شده است [۱۸] . با اتمام تناظریابی، تمام داده‌های مرجع و مردم‌گستر به صورت دارای متناظر و بدون متناظر دسته‌بندی می‌شوند تا در مراحل بعدی، آنالیز کیفیت روی آن‌ها صورت گیرد.

### ۳-۱- محاسبه دقت توصیفی

داده‌های مکانی اغلب اوقات دارای اقلام توصیفی[۱] هستند که حاوی اطلاعاتی افزون بر مکان داده‌اند. اقلام توصیفی ممکن است شامل نام عارضه، طول یا عرض آن، کاربری یک عارضه و یا اطلاعاتی از این دست باشند. مقیاسهای اندازه گیری اقلام توصیفی را می توان در چهار دسته تقسیم‌بندی کرد: اسمی[۲]، ترتیبی[۳]، نسبی[۴] و بازه‌ای[۵] [۱۹]. برای ارزیابی دقت اطلاعات توصیفی، بر اساس نوع قلم توصیفی (یا ویژگی) و اینکه در کدام یک از دسته‌های بالا قرار می‌گیرد، روش‌های مختلفی وجود دارد. داده‌های نسبی و بازه‌ای از آنجایی که ماهیت عددی دارند و به صورت یک کمیت بیان می‌شوند، به راحتی قابل مقایسه با یکدیگرند و بنابراین می‌توان دقت آن‌ها را به سادگی و با مقایسه به دست آورد. داده‌های ترتیبی نیز معمولاً دامنه مقادیر قابل قبول کوچکی دارند. مثلاً نوع کاربری یک عارضه معمولاً یک مقدار از مجموعه‌ای از مقادیر از پیش تعیین شده می‌گیرد. بنابراین ارزیابی کیفیت این نوع از اطلاعات توصیفی نیز نسبتاً ساده است و برای ارزیابی آن‌ها معمولاً از روش‌های طبقه‌بندی استفاده می‌شود [۲۰]. اما در مورد داده‌های اسمی روش‌های ارزیابی دقت سخت تر و پیچیده ترند. چرا که این گونه از داده‌ها به صورت یک نام یا در واقع یک رشته از حروف بیان می‌شوند. در صورتی که دو رشته مورد بررسی (مثلاً یک نام از مجموعه‌داده مرجع و یک نام از مجموعه مردم‌گستر)

---

۱ Attribute
۲ Nominal
۳ Ordinal
٤ Ratio
٥ Interval

دقیقا مشابه هم باشند، می‌توان با اطمینان از دقت اطلاعات سخن به میان آورد اما چنانچه با هم اختلاف داشته باشند، نمی‌توان به سادگی یکی از آن‌ها را اشتباه فرض کرد. چراکه ممکن است دو رشته تنها در یک حرف با هم فرق داشته باشند که در این صورت باز هم دقت داده مورد نظر می تواند قابل قبول باشد.

در بسیاری از موارد، داده‌های مردم‌گستر فاقد اطلاعات توصیفی کافی هستند. چرا که تولید کنندگان این داده‌ها، که مردم عادی هستند، دانش کافی در مورد اطلاعات توصیفی مربوط به عوارض ندارند و یا اطلاعات آن‌ها اشتباه است. مثلاً ممکن است تعداد زیادی از کاربران در تشخیص نوع یک عارضه‌ی راه (شریانی، درجه ۱، دسترسی و غیره) اشتباه کنند و برداشت آن‌ها از نوع راه با تعاریف رسمی تفاوت زیادی داشته باشد. در بسیاری از اوقات هم مشارکت کنندگان تنها به وارد کردن نام داده بسنده کرده و به دنبال تکمیل سایر اطلاعات توصیفی عوارض نیستند. بنابراین اطلاعات توصیفی مربوط به عوارض مردم‌گستر در بسیاری از موارد یا ناقص است و یا اشتباه؛ به ویژه در مورد اقلام توصیفی غیر از نام عارضه. این واقعیت به همراه دلیل دیگری که پیش‌تر به آن اشاره شد، یعنی سادگی محاسبه اقلام توصیفی نسبی، بازه‌ای و ترتیبی نسبت به اسمی، باعث گردید تا در این تحقیق تنها به بررسی کیفیت اطلاعات توصیفی از نوع اسمی پرداخته شود .

مجموعه‌داده مردم‌گستر مورد استفاده در این تحقیق (یعنی داده‌های سایت OpenStreetMap) شامل اقلام توصیفی مختلفی است که عبارتند از : شماره عارضه (id)، نوع عارضه (راه درجه ۱، درجه ۲ و غیره) مسیر تردد (یک طرفه بودن یا نبودن)، حداکثر سرعت، و اینکه عارضه مورد بحث پل یا تونل است. در شکل ۱ جدول توصیفات مربوط به این داده‌ها آمده است. در این تحقیق تنها به بررسی دقت نام عارضه پرداخته می‌شود.

Table

roads

| FID | Shape * | osm_id | name | ref | type | oneway | bridge | tunnel | maxspeed |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Polyline | 4292665 | قاسمی | | residential | 0 | 0 | 0 | 0 |
| 1 | Polyline | 4292748 | تیموری | | secondary | 1 | 0 | 0 | 0 |
| 2 | Polyline | 4292765 | صالحی | | secondary_link | 1 | 0 | 0 | 0 |
| 3 | Polyline | 4292778 | شادمهر | | secondary | 0 | 0 | 0 | 0 |
| 4 | Polyline | 4292779 | زنجان شمالی | | secondary | 1 | 0 | 0 | 0 |
| 5 | Polyline | 4292780 | بهبوی | | secondary | 0 | 0 | 0 | 0 |
| 6 | Polyline | 4292781 | Habibollah Bldv. | | secondary | 0 | 0 | 0 | 0 |
| 7 | Polyline | 4292787 | ولدخانی | | residential | 0 | 0 | 0 | 0 |
| 8 | Polyline | 4292788 | استوار | | tertiary | 0 | 0 | 0 | 0 |
| 9 | Polyline | 4292854 | یکم دریان‌نو | | tertiary | 0 | 0 | 0 | 0 |
| 10 | Polyline | 4292857 | دوازدهم دریان‌نو | | residential | 0 | 0 | 0 | 0 |
| 11 | Polyline | 4292858 | دهم دریان‌نو | | residential | 0 | 0 | 0 | 0 |
| 12 | Polyline | 4292879 | خوشرو | | residential | 0 | 0 | 0 | 0 |
| 13 | Polyline | 4292880 | راستان | | residential | 0 | 0 | 0 | 0 |
| 14 | Polyline | 4293070 | حبیب‌زادگان | | tertiary | 0 | 0 | 0 | 0 |
| 15 | Polyline | 4294716 | | | secondary | 1 | 0 | 0 | 0 |
| 16 | Polyline | 4294733 | اقبال آشتیانی | | residential | 0 | 0 | 0 | 0 |
| 17 | Polyline | 4294735 | زوالفقاری | | residential | 0 | 0 | 0 | 0 |
| 18 | Polyline | 4294743 | جناح | | trunk | 1 | 0 | 0 | 0 |
| 19 | Polyline | 4294746 | Shahid Saremi Street | | secondary | 0 | 0 | 0 | 0 |
| 20 | Polyline | 4294833 | نیکروش فرد | | residential | 0 | 0 | 0 | 0 |

2 (0 out of 217666 Selected)

roads

شکل ۱- جدول اقلام توصیفی داده‌های مردم‌گستر (سایت OpenStreetMap)

برای ارزیابی دقت توصیفی می‌بایست اطلاعات توصیفی داده‌های مردم‌گستر را با اطلاعات توصیفی داده‌های مرجع مقایسه کرد. برای این منظور، علاوه بر عوارض دارای متناظر، عوارض بدون متناظر هم مورد بررسی قرار می‌گیرند. چرا که ممکن است یک عارضه دو نمایش متفاوت در دو مجموعه‌داده داشته باشد و این دو نمایش به قدری از هم اختلاف مکانی (هندسی) داشته باشند که در مرحله تناظریابی به عنوان عوارض غیر نظیر طبقه‌بندی شوند. اما نام و در حالت کلی تر، اطلاعات توصیفی آن‌ها با هم یکسان باشد.

برای مقایسه دو رشته، توابع و روش‌های مختلفی وجود دارد همچون روش Metaphone که از قواعد تلفظ زبان انگلیسی برای مقایسه استفاده می‌کند و یا الگوریتم Levenshtein که فاصله بین دو رشته را بر اساس تعداد

ویرایش‌های لازم برای تبدیل یک رشته به رشته دیگر اندازه‌گیری می‌کند. یک روش پر کاربرد دیگر روش مشابهت متن[1] است که میزان مشابهت دو رشته به یکدیگر را بر اساس طول رشته مشخص می‌کند [۲۱]. از میان این روش‌ها، الگوریتم Levenshtein در مباحث مربوط به مقایسه و تناظریابی بین دو رشته پرکاربردتر از بقیه است و دقت بالاتری دارد [۱۳]. ضمن اینکه حالت کلی تری از بقیه دارد و برای سایر زبان‌ها غیر از زبان انگلیسی هم به راحتی و بدون نیاز به تغییر قابل استفاده است. بنابراین، در این تحقیق از این الگوریتم برای محاسبه دقت اطلاعات توصیفی استفاده شده است اما برای بهبود کارایی و دقت آن مجموعه ای از پیش پردازش‌های متنی با استفاده از دانش محلی طراحی شده اند. ضمنا بر خلاف سایر تحقیقاتی که از این الگوریتم استفاده کرده اند، در این مقاله برای محاسبه دقت نام یک عارضه، طول نام آن هم مدنظر قرار گرفته است.

الگوریتم Levenshtein تعداد ویرایش‌های لازم برای تبدیل یک رشته به رشته دیگر را اندازه‌گیری کرده و به عنوان یک فاصله بیان می‌کند (به همین دلیل نام دیگر آن فاصله‌ی ویرایش[2] است). این ویرایش‌ها شامل سه عملگر افزودن یک حرف (کاراکتر)، حذف یک حرف و یا جابجایی بین دو حرف می‌باشد [۲۲].

اگر دو رشته $a$ و $b$ در نظر گرفته شوند، تابع فاصله بین این دو که با $lev_{a,b}(|a|,|b|)$ نمایش داده می‌شود به صورت زیر خواهد بود [۲۲]:

(۱)

$$llev_{a,b}(i,j) = \begin{cases} \max(i,j) & if\ \min(i,j)=0 \\ \min\begin{cases} lev_{a,b}(i-1,j) \\ lev_{a,b}(i,j-1) \\ lev_{a,b}(i-1,j-1)+1(a_i \neq b_j) \end{cases} & otherwise \end{cases}$$

که در آن $i$ و $j$ به ترتیب برابر طول رشته $a$ و $b$ است و $1(a_i \neq b_j)$ تابع علامت است و زمانی که $a_i = b_j$ برابر صفر و در بقیه حالات برابر ۱ خواهد بود. حداقل مقدار فاصله ویرایش بین دو رشته برابر تفاضل طول آن دو رشته از هم، و حداکثر مقدار آن برابر طول رشته‌ی بلندتر است. به طور مثال فاصله بین دو رشته "سعادت" و "مسعود" طبق این الگوریتم برابر ۳ است. چرا که برای رسیدن از عبارت سعادت به مسعود، سه ویرایش به شرح زیر نیاز است:

- سعادت ← مسعادت (افزودن حرف م)
- مسعادت ← مسعودت (تبدیل حرف الف به واو)
- مسعودت ← مسعود (حذف حرف ت)

| | س | ع | ا | د | ت |
|---|---|---|---|---|---|
| + | = | = | ↕ | = | - |
| م | س | ع | و | د | |

پیش از شروع بررسی دقت توصیفی می‌بایست با اعمال یک سری پیش پردازش متنی، نام عوارض را در دو مجموعه داده استاندارد کرد. این پیش پردازش‌ها با از بین بردن اختلافات جزئی که در بیان نام عوارض در دو مجموعه موجود است، و بدون تغییر دادن هسته اصلی نام عارضه، باعث بالا رفتن دقت الگوریتم Levenshtein می‌شوند. این پیش پردازش‌ها عبارتند از:

۱- کلمه‌ها و یا اختصاراتی که نشان دهنده نوع راه هستند مثل کوچه، خیابان، بلوار، "ک." و "خ." حذف شده و در نظر گرفته نمی‌شوند. این کار دو دلیل دارد:

الف) ممکن است در تعبیر نوع یک راه بین تولید کننده رسمی و مردم عادی اختلاف نظر وجود داشته باشد. مثلاً فرض کنید نام یک عارضه در مجموعه مرجع "کوچه مطلبی" و در مجموعه مردم‌گستر "خیابان مطلبی" باشد. در حالت عادی فاصله ویرایشی بین این دو رشته بزرگ‌تر از حد قابل قبول است و بنابراین، این دو به عنوان عوارض نامتناظر طبقه‌بندی می‌شوند. در حالی که در حقیقت نام عارضه درست است. با حذف عباراتی مثل کوچه و بن بست، این دو رشته با هم یکسان خواهند شد.

ب) هم در مجموعه مرجع و هم مردم‌گستر، گاهی اوقات به جای به کار بردن کلماتی همچون "خیابان" و "کوچه" از اختصاراتی همچون "خ." و "ک." استفاده شده است. با حذف چنین کلماتی، دقت محاسبه دقت توصیفی بالاتر می‌رود.

۲- کلمه "شهید" هم با دلایلی مشابه با دلایل حالت قبل از نام عوارض حذف می‌شود.

۳- در برخی از مناطق مورد بررسی این تحقیق، برای نام‌گذاری خیابان‌های فرعی یا کوچه‌ها از نام خیابان اصلی به همراه یک عدد استفاده می‌شود (شکل ۲).

[1] Text similarity
[2] Edit distance


شکل ۲- نمونه‌ای از معابر که نام آن‌ها از یک رشته حرف به علاوه یک عدد تشکیل شده اند

همانگونه که در شکل ۲ مشاهده می‌شود، در یک محدوده جغرافیایی کوچک تعداد قابل توجهی خیابان فرعی با نام متشکل از یک رشته و یک عدد وجود دارد. از آنجایی که این نام‌ها تنها در یک کاراکتر با هم فرق دارند، فاصله ویرایش آن‌ها ۱ است و ممکن است در الگوریتم طراحی شده به اشتباه متناظر با هم در نظر گرفته شوند؛ البته در چنین حالتی مجاورت مکانی آن‌ها نیز مزید بر علت شده و باعث گمراهی بیش‌تر الگوریتم می‌شود. ضمن اینکه نحوه نام‌گذاری آن‌ها در دو مجموعه معمولاً متفاوت است مثلاً نام یک عارضه در یک مجموعه، به صورت "عطایی ۱" و در مجموعه دیگر به صورت "عطایی اول" ذخیره شده است. برای رفع این مشکل، تمام اعداد موجود در نام عوارض به معادل حروفی خود تبدیل می‌شوند (مثلاً ۱ به "اول" و "یکم" تبدیل می‌شود و هر دو حالت در نظر گرفته می‌شود).

۴- برخی از نام‌ها در مجموعه مردم‌گستر با حروف لاتین ذخیره شده اند[1]. بنابراین مقایسه آن‌ها با نام‌های مجموعه مرجع نتیجه غلطی تولید می‌کند. برای رفع این مشکل دو رویکرد مختلف وجود دارد: ۱) حذف تمام نام‌هایی که به صورت لاتین ذخیره شده اند ۲) تبدیل این رشته‌ها به فارسی. از آنجایی که رویکرد اول باعث از بین رفتن بخشی از داده‌های مردم‌گستر می‌شود، در این تحقیق از رویکرد دوم استفاده شده است. بنابراین قبل از شروع محاسبه دقت توصیفی، تمام نام‌هایی که با حروف لاتین ذخیره شده اند با استفاده از توابع تبدیل به رشته‌هایی با حروف فارسی تبدیل می‌شوند.

با در نظر گرفتن این نکات، نوبت به محاسبه دقت توصیفی می‌رسد. این فرآیند از دو مرحله کلی تشکیل می‌شود: محاسبه دقت عوارض دارای متناظر و محاسبه دقت عوارض بدون متناظر.

## ۳-۲- محاسبه دقت توصیفی عوارض دارای متناظر

ابتدا نام هر یک از عوارض مرجع دارای متناظر با نام عارضه مردم‌گستر متناظر با آن مقایسه شده و فاصله ویرایش بین دو رشته با استفاده از الگوریتم Levenshtein محاسبه می‌شود. فاصله دو رشته از هم می‌تواند مقداری بین صفر تا بیشینه طول بین دو رشته داشته باشد. اگر این فاصله برابر صفر باشد، یعنی دو رشته یکسان‌اند و بنابراین نام عارضه مردم‌گستر صحیح است. اگر فاصله برابر ۱ یا ۲ باشد، می‌توان تفاوت بین دو رشته را به اشتباهات تایپی یا سهوی در هنگام ورود داده توسط کاربران نسبت داد. هرچند چنین اظهارنظری نیاز به بررسی بیش‌تری دارد. اما فواصل مساوی یا بزرگ‌تر از ۳ حاکی از دو رشته

[1] فینگلیش

کاملا متفاوت است و اگر فاصله بین دو رشته برابر این مقدار باشد یعنی اطلاعات توصیفی (یا در واقع نام) وارد شده اشتباه است. در جدول ۱ می‌توان مقادیر مختلف ممکن برای فاصله ویرایش و تعبیر هر یک از این مقادیر را مشاهده کرد:

جدول ۱- مقادیر فاصله ویرایش ممکن بین دو رشته و تعبیر هر یک از آن‌ها

| فاصله بین دو رشته (d) | تعبیر فاصله |
|---|---|
| ۰ | دو رشته یکسان‌اند |
| ۱ | دو رشته تنها در یک کاراکتر اختلاف دارند که احتمالاً ناشی از اشتباهات سهوی است |
| ۲ | دو رشته در دو کاراکتر اختلاف دارند. نیاز به بررسی بیشتر است |
| $\geq 3$ | دو رشته کاملا متفاوت از همدیگرند |

برای حالتی که فاصله بین دو رشته ۱ یا ۲ واحد است، مقدار فاصله بر کمینه طول بین دو رشته تقسیم می‌شود. چنانچه حاصل کمتر از ۰/۴ باشد، اختلاف قابل قبول بوده و دو رشته با هم برابر در نظر گرفته می‌شوند و در غیر این صورت دو رشته متفاوت در نظر گرفته می‌شوند. مقدار ۰/۴ با استفاده از سعی و خطا و با آزمودن مقادیر مختلف برای حد قابل قبول به دست آمده است.

بنابراین در این مرحله، فاصله‌ی ویرایش نام هر عارضه مردم‌گستر دارای متناظر (البته در صورتی که عارضه دارای قلم توصیفی نام باشد)، با نام عارضه مرجع نظیر با آن محاسبه می‌شود. اگر این فاصله برابر صفر باشد، یعنی نام عارضه مردم‌گستر صحیح است و در جدول اطلاعات توصیفی مربوط به عارضه و در یک فیلد جدید با نام AAC[1] مقدار $a$ برای آن عارضه ذخیره می‌شود. اگر فاصله ویرایش برابر ۱ یا ۲ باشد، نسبت این فاصله به کمینه طول دو رشته محاسبه شده و چنانچه این مقدار کمتر از ۰/۴ باشد، در فیلد AAC مقدار $b$ ذخیره می‌شود. اگر نسبت محاسبه شده بیش‌تر از ۰/۴ باشد یا فاصله ویرایش بین دو رشته بزرگ‌تر یا مساوی ۳ باشد، در فیلد AAC مقدار $c$ ذخیره می‌شود (جدول ۲).

## ۳-۳- محاسبه دقت توصیفی عوارض بدون متناظر

در این مرحله عوارض مردم‌گستری که دارای قلم توصیفی نام بوده و متناظری در مجموعه مرجع ندارند مورد بررسی قرار می‌گیرند. برای هر کدام از این عوارض، یک فاصله جستجو بر اساس رابطه زیر در نظر گرفته می‌شود [۲۱]:

$$D_s = A + \frac{W}{2} \qquad (۲)$$

در این رابطه $A$ بیانگر دقت مختصات مسطحاتی دستگاه‌های تعیین موقعیت GPS[2] مورد استفاده توسط عموم است؛ که به طور معمول برابر ۵ متر در نظر گرفته می‌شود [۲۳]. اما در اینجا برای در نظر گرفتن بدترین حالت ممکن، ۱۵ متر در نظر گرفته شده است. $W$ هم برابر عرض راه (عارضه) مورد بررسی است و برای این در نظر گرفته شده است که خطاهای ناشی از برداشت داده در کناره‌های راه به جای وسط راه را پوشش دهد.

سپس نام عوارض مرجعی که درون این فاصله جستجو قرار دارند مورد بررسی قرار گرفته و فاصله‌ی ویرایش هر کدام از آن‌ها با نام عارضه مردم‌گستر مورد بررسی محاسبه می‌شود. اگر عارضه‌ی مرجعی با فاصله ویرایش صفر درون فاصله جستجو پیدا شود، آن عارضه به عنوان متناظرِ عارضه مردم‌گستر مورد بررسی در نظر گرفته شده و هم زمان در فیلد AAC، برای عارضه مردم‌گستر مقدار a ذخیره می‌شود. نکته قابل توجه این است که در این مرحله برای بالا بردن اطمینان تنها فاصله ویرایش صفر به عنوان مقدار مطلوب در نظر گرفته می‌شود و هر مقداری غیر از آن غیر قابل قبول تلقی می‌شود.

با اتمام مراحل فوق فیلد AAC برای هر عارضه مردم‌گستر دارای یک مقدار می‌شود. بر اساس این مقادیر می‌توان داده‌ها را در ۴ دسته تقسیم‌بندی کرد:

[1] Attribute Accuracy Condition

[2] Global Positioning System

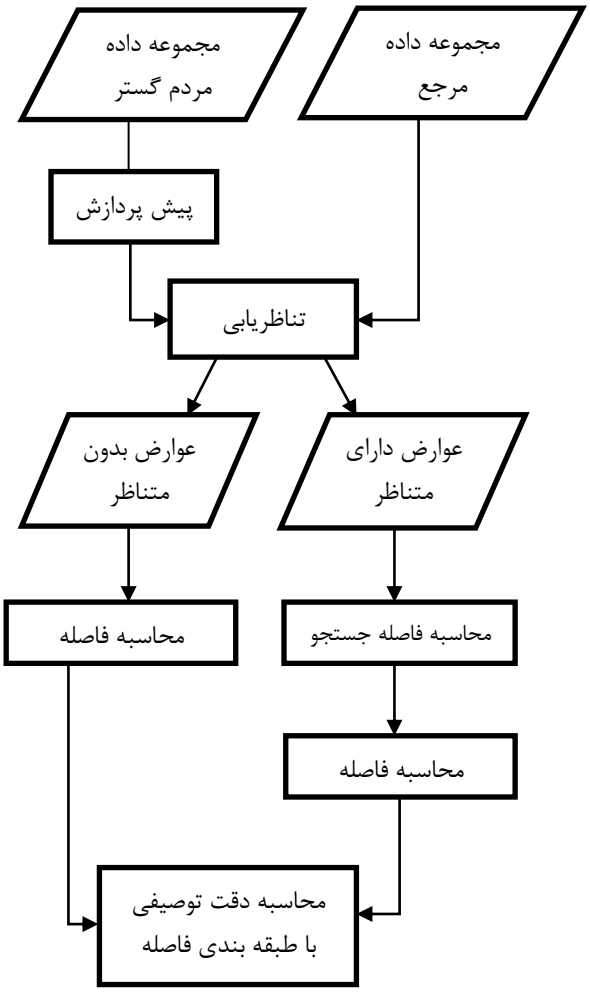
۱- داده‌هایی که دارای مقدار $a$ هستند و نام آن‌ها دقیق است

۲- داده‌هایی که دارای مقدار $b$ هستند و نام آن‌ها تقریباً دقیق است

۳- داده‌های دارای مقدار $c$ که نام آن‌ها اشتباه است

۴- داده‌های بدون مقدار که فاقد اطلاعات توصیفی (نام) هستند.

با تقسیم مجموع طول داده‌های دسته اول و دوم بر مجموع طول داده‌های مردم‌گستری که دارای نام هستند، می‌توان دقت توصیفی داده‌های مردم‌گستر را محاسبه کرد. در ضمن با تقسیم مجموع طول داده‌های هر یک از دسته‌های فوق بر مجموع طول داده‌های دارای نام و نیز مجموع طول کل داده‌ها می‌توان به برآورد کامل‌تری از دقت توصیفی رسید [۱۸]. مراحل روش پیشنهادی را در شکل روبرو مشاهده می کنید.




شکل ۳- فلوچارت مراحل کلی کار

برای ارائه دقت برآورد شده روش‌های مختلفی وجود دارد. به طور مثال می‌توان میزان دقت را روی یک نقشه و با استفاده از متغیرهای بصری نمایش داد و یا می‌توان مقادیر دقت را به صورت کمی و در قالب یک جدول گزارش کرد. طبیعتاً ارائه بصری درک بهتری از دقت به کاربر می‌دهد. این مزیت برای داده‌های مردم‌گستر که بخش زیادی از کاربران آن مردم عادی هستند اهمیت بیش‌تری پیدا می‌کند چراکه اگر کاربر درک بهتری از کیفیت داده داشته باشد، بهتر می‌تواند در استفاده از آن داده تصمیم‌گیری کند. بنابراین در این تحقیق برای ارائه دقت توصیفی از متغیر بصری "رنگ" استفاده می‌شود [۲۴]. بدین صورت که عوارض بر اساس مقدار عددی دقت آن‌ها به ۳ دسته یا کلاس‌های مختلف تقسیم شده و عوارض مربوط به هر کلاس با استفاده از یک رنگ مشخص نمایش داده می‌شوند.

مزیت اصلی روش پیشنهادی این تحقیق نسبت به سایر روش های استفاده شده در برآورد دقت توصیفی (و یا دقت اطلاعات متنی) استفاده از الگوریتم Levenshtein و پیش پردازش‌های صورت گرفته برای مقایسه دو رشته است. این الگوریتم نسبت به سایر الگوریتم های مشابه برای مقایسه دو رشته که در تحقیقات مربوط به اطلاعات مردم گستر استفاده شده ([۱۰][۱۳][۱۴]) دقت بالاتری داشته است. بزرگترین نقطه ضعف این الگوریتم در مورد حروف مخفف است [۱۴]؛ که این مشکل در این تحقیق با در نظر گرفتن قیود مختلف از بین رفته است. همچنین در این تحقیق با بررسی عوارض بدون متناظر، علاوه بر عوارض دارای متناظر، برآورد کاملی از دقت توصیفی ارائه شده است و همزمان تمامیت توصیفی هم محاسبه شده است.

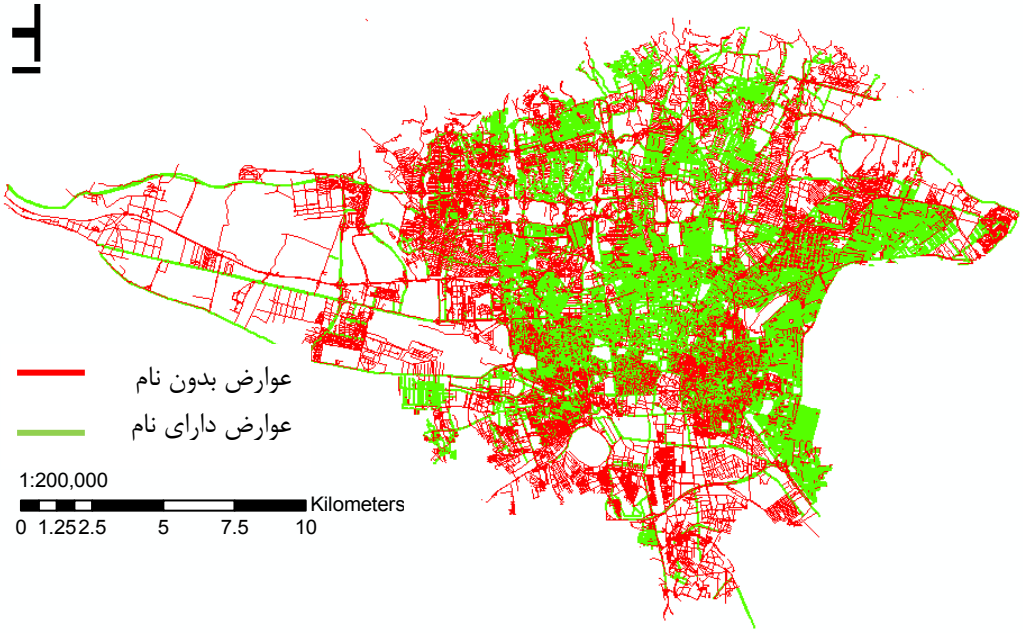
## ۴- پیاده سازی و ارزیابی روش

منطقه مورد مطالعه در این تحقیق، شهر تهران می‌باشد. این شهر با مساحت تقریبی ۷۳۰ کیلومتر مربع و با جمعیتی حدود ۸ میلیون نفر در مرکز ایران واقع شده و پایتخت کشور است [۲۵]. البته در این مطالعه مناطق حومه تهران مورد بررسی قرار نگرفته و تنها مناطق ۲۲ گانه شهر با مساحت ۵۹۳ کیلومتر مربع بررسی شده اند. داده‌های شبکه راه‌ها و معابر تهران، تولیدی شهرداری

تهران، با مقیاس ۱:۲۰۰۰ به عنوان مجموعه‌داده مرجع، و داده‌های سایت OpenStreetMap (OSM) به عنوان مجموعه‌داده مردم‌گستر مورد استفاده این تحقیق قرار گرفتند. جدول ۲ طول کل عوارض مردم‌گستر و طول عوارض دارای نام و بدون نام را در این مجموعه نشان داده و شکل ۴ نقشه این دو دسته عارضه (عوارض دارای نام و بدون نام) را نمایش می دهد.

جدول ۲- طول عوارض مردم‌گستر دارای نام و بدون نام

| طول کل عوارض مردم‌گستر(متر) | طول عوارض بدون نام (متر) | طول عوارض دارای نام (متر) |
|---|---|---|
| ۸۳۶۶۶۹۳ | ۴۴۷۰۹۱۱ (٪ ۵۳) | ۳۸۹۵۷۸۳ (٪ ۴۷) |



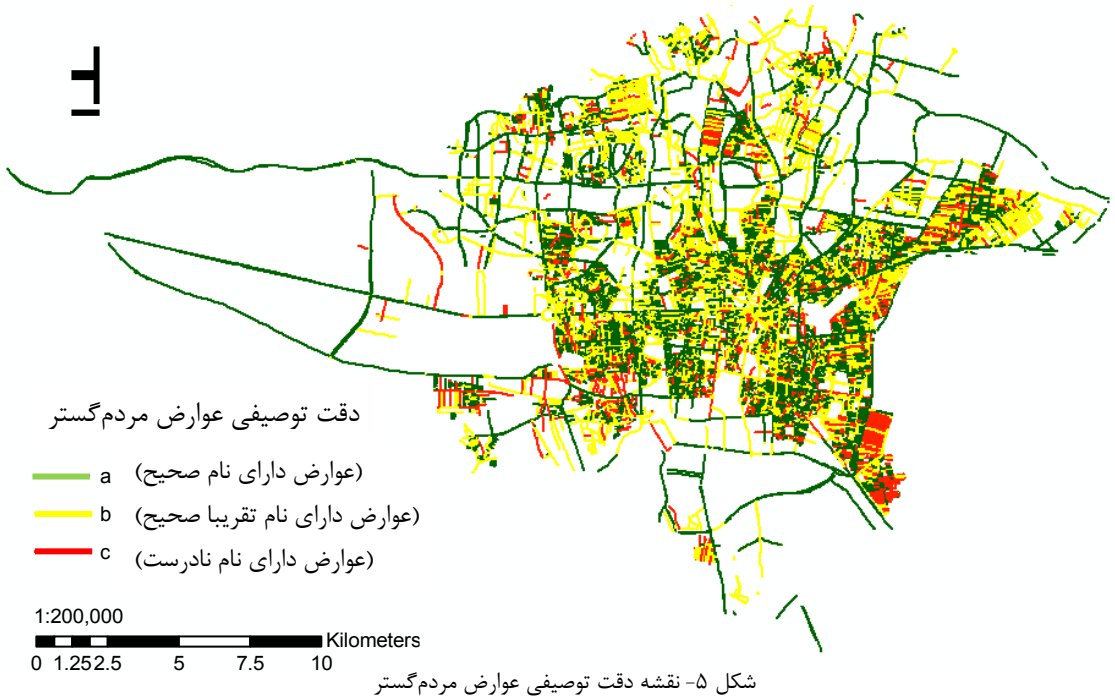

شکل ۴- نقشه عوارض مردم‌گستر دارای نام و بدون نام

شکل ۵ وضعیت دقت نام عوارض مردم‌گستر را بر اساس فاصله ویرایش (و مقدار فیلد AAC) نشان می‌دهد. در این شکل عوارض دارای مقدار a به رنگ سبز، مقدار b به رنگ زرد و مقدار c به رنگ قرمز نمایش داده شده اند.

همچنین جدول ۳ مجموع طول عوارض هر کدام از ۳ دسته فوق و نسبت طول هر کدام از دسته‌ها را به طول کل عوارض مردم‌گستر دارای نام نشان می‌دهد.

جدول ۳- طول عوارض مردم‌گستر بر اساس دقت نام آن‌ها

| مقدار فیلد *AAC* | مجموع طول عوارض دسته (متر) | نسبت طول عوارض دسته به طول کل عوارض دارای نام |
|---|---|---|
| *a* | ۱۲۷۳۹۲۲ | ٪ ۳۲/۷ |
| *b* | ۱۷۴۱۴۱۵ | ٪ ۴۴/۷ |
| *c* | ۸۸۰۴۴۶ | ٪ ۲۲/۶ |




شکل ۵- نقشه دقت توصیفی عوارض مردم‌گستر

همان طور که از جدول بالا و شکل ۵ پیداست، عوارض مردم‌گستر از دقت توصیفی نسبتاً بالایی برخوردار هستند. نام ۳۳ درصد از این عوارض با نام رسمی آن‌ها دقیقا یکسان است و حدود ۴۵ درصد از آن‌ها نام تقریباً صحیح دارند. البته دلیل بالا بودن طول عوارض دسته اول (مقدار a برای فیلد AAC) این است که اکثر بزرگراه‌ها و راه‌های اصلی، که عموماً عوارضی با طول‌های بلند هستند، در این دسته جای می‌گیرند. چراکه این عوارض، عوارض شناخته شده و مهمی هستند و بنابراین نام آن‌ها صحیح وارد شده است. این نکته با دقت در شکل ۵ هم پیداست؛ اکثر بزرگراه‌ها در این نقشه به رنگ سبز هستند یعنی نام آن‌ها صحیح است.

برای محاسبه دقت توصیفی کل مجموعه مردم‌گستر می‌بایست مجموع طول عوارض دارای نام صحیح ($AAC=a$) و یا تقریباً صحیح ($AAC=b$) را بر مجموع طول کل عوارض دارای نام ($AAC= a$ یا $b$ یا $c$) تقسیم کرد. لازم به ذکر است که در نظر گرفتن کل عوارض مردم‌گستر (به جای آن‌هایی که دارای نام هستند) به عنوان مخرج کسر، برآورد غلطی از دقت توصیفی تولید می‌کند چرا که هدف در این قسمت این است که مشخص شود چه تعدادی یا در واقع چه درصدی از توصیفات تولید شده توسط داوطلبان صحیح هستند.

بنا بر آنچه ذکر شد دقت توصیفی از رابطه ۳ بدست می آید. با تقسیم مجموع طول عوارض مردم‌گستر دارای نام بر مجموع طول کل عوارض مردم‌گستر هم می‌توان به درصد تمامیت داده‌های مردم‌گستر از نظر دقت توصیفی رسید (رابطه ۴).

$$\text{دقت توصیفی کل داده‌های مردم‌گستر} = \frac{\text{مجموع طول عوارض مردم گستر دارای نام صحیح و یا قابل قبول}}{\text{مجموع طول کل عوارض مردم‌گستر دارای نام}} \times 100 = \%77.4 \quad (۳)$$

$$\text{تمامیت توصیفی داده‌های مردم‌گستر} = \frac{\text{مجموع طول عوارض مردم گستر دارای نام}}{\text{مجموع طول کل عوارض مردم‌گستر}} \times 100 = \%47 \quad (۴)$$

## ۴-۱- ارزیابی

برای ارزیابی روش ارائه شده در این تحقیق رایج‌ترین راهکار موجود مشخص کردن مناطق مختلفی از منطقه مورد مطالعه (به عنوان مناطق تست یا کنترل) به صورت تصادفی، ارزیابی دستی دقت در آن مناطق و مقایسه نتایج به دست آمده با نتایج حاصل از روش استفاده شده است [۱۰][۱۲][۱۴].

در این تحقیق برای مشخص کردن نقاط تست ابتدا ۱۰۰ نقطه با مختصات تصادفی درون منطقه مورد مطالعه تولید می‌شوند. سپس حول هر کدام از این نقاط یک دایره (دایره تست) به مساحت ۱ کیلومتر مربع (شعاع ۵۶۴ متر)

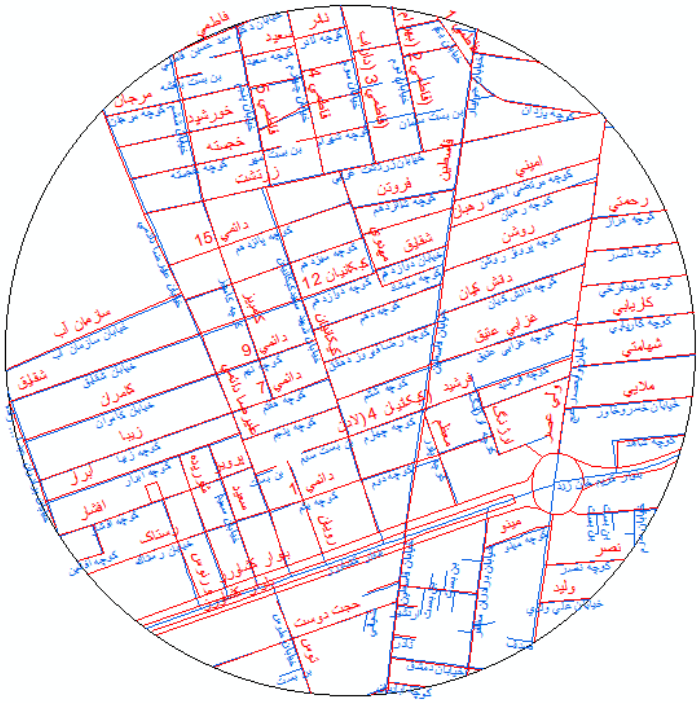
ایجاد شده و عملکرد روش محاسبه دقت توصیفی درون این دایره‌ها به صورت دستی مورد بررسی قرار می‌گیرد. به این ترتیب در مجموع منطقه‌ای به مساحت ۱۰۰ کیلومتر مربع پوشش داده می‌شود که تقریباً برابر با یک ششم (حدوداً برابر با ۱۶ درصد) مساحت کل منطقه مورد مطالعه است.

برای ارزیابی دقت توصیفی، نام عوارض مردم‌گستر واقع درون دایره‌های تست به صورت دستی با نام عوارض مرجع واقع در این دایره‌ها مقایسه شده و عوارض دارای نام صحیح، نام تقریباً صحیح و نام نادرست مشخص می‌گردند. سپس مجموع طول این عوارض با مجموع طول حاصل از روش پیشنهادی، که در قسمت قبل محاسبه شد، مقایسه می‌شود. شکل شماره ۶ یکی از ۱۰۰ دایره تست و عوارض درون آن را نشان می‌دهد. عوارض مردم‌گستر در این شکل به رنگ قرمز و عوارض مرجع به رنگ آبی دیده می‌شوند.

نتایج حاصل از ارزیابی در جدول ۴ آمده است.

جدول ۴- نتایج ارزیابی دقت توصیفی

| روش محاسبه | طول عوارض دارای نام | طول عوارض دارای نام صحیح | طول عوارض دارای نام تقریباً صحیح | طول عوارض دارای نام نادرست |
|---|---|---|---|---|
| **خودکار (متر)** | ۷۴۳۶۰۸ | ۳۰۴۸۷۹ | ۲۹۰۰۰۷ | ۱۴۸۷۲۲ |
| **دستی (متر)** | ۷۴۳۶۰۸ | ۳۳۸۴۱۵ | ۳۵۹۶۰۸ | ۴۵۵۸۵ |
| **درصد اختلاف** | ۰ | + ٪ ۱۱ | + ٪ ۲۴ | - ٪ ۳۰ |



شکل ۶- نام عوارض مرجع و مردم‌گستر درون یک دایره تست

در این جدول ستون سوم بیانگر عوارض دارای مقدار a برای فیلد AAC است؛ ستون چهارم و پنجم هم به ترتیب بیانگر عوارض دارای مقدار b و c هستند. سطر آخر این جدول در واقع نشان دهنده دقت روش استفاده شده است که در مورد عوارض دارای نام صحیح ۸۹ درصد، در مورد عوارض دارای نام تقریبا صحیح ۷۶ درصد و در مورد عوارض دارای نام نادرست ۷۰ درصد است و میانگین آن در حالت کلی برابر ۷۸ درصد است.

دلیل اختلاف موجود بین عملکرد این روش با روش محاسبه دستی، پیچیدگی‌ها و ظرافت‌های زبانی است که قدرت هوش مصنوعی در درک آن‌ها بسیار کمتر از هوش انسانی است. در واقع بررسی صحت یک عبارت یا کلمه

نیاز به هوش انسانی دارد. به طور مثال، اگر نام یک عارضه در مجموعه مرجع "کوچه محمد عرب" و در مجموعه مردم‌گستر "کوچه عرب" باشد، الگوریتم طراحی شده، نام عارضه مردم‌گستر را به عنوان نادرست طبقه‌بندی می‌کند. حال آنکه هر انسانی این نام را صحیح تشخیص می‌دهد.

## ۵- نتیجه گیری

همانگونه که پیشتر اشاره شد، هدف این تحقیق بررسی دقت توصیفی اطلاعات مکانی مردم‌گستر بود. داده‌های مردم‌گستر اقلام توصیفی مختلفی همچون "نام عارضه" و "نوع عارضه (راه)" دارند که در بخش ۳ به آن‌ها اشاره شد. اما از آنجا که به جز نام، مقدار سایر اقلام توصیفی برای درصد قابل توجهی از داده‌ها تهی است و یا دقت خوبی ندارد، در این تحقیق فقط به بررسی نام عوارض پرداخته شد و دقت توصیفی در قالب دقت نام عوارض محاسبه گردید.

برای بررسی دقت نام عوارض مردم‌گستر، اختلاف بین نام این عوارض با نام عوارض مرجعِ نظیر، از طریق الگوریتم Levenshtein محاسبه شد. برای بهبود کارایی این الگوریتم و از بین بردن نقاط ضعف آن از مجموعه‌ای از پیش پردازش‌های متنی که بر اساس دانش محلی به دست آمده اند استفاده شد. این پیش پردازش‌ها باعث استاندارد شدن نام عوارض شده و با از بین بردن مواردی همچون کلمات زائد باعث بالا رفتن دقت روش مورد استفاده می‌شوند.

همچنین برای بهبود بیشتر این الگوریتم، برای مقایسه بین دو رشته، طول آن دو نیز مد نظر قرار گرفته و میزان اختلاف به صورت نسبی بررسی می‌شود. این در حالی است که در سایر تحقیقات مشابه [۱۲و۱۳] تنها یک عدد ثابت (معمولا ۳ یا ۴) به عنوان حد آستانه قابل قبول برای میزان اختلاف بین دو رشته در نظر گرفته شده و طول رشته مورد بررسی نادیده گرفته شده است. در صورتی که بدیهی است که هر چه طول یک رشته کوتاهتر باشد، میزان خطای قابل قبول برای آن هم می‌بایست کوچکتر باشد.

این اختلاف که فاصله ویرایش نام دارد می‌تواند مقداری بین صفر تا تعداد کاراکترهای عارضه مورد بررسی داشته باشد. پس از آن عوارض مردم‌گستر دارای نام بر اساس فاصله ویرایش محاسبه شده، در سه دسته طبقه‌بندی شدند: عوارض دارای نام صحیح، عوارض دارای نام تقریباً صحیح و عوارض دارای نام ناصحیح. طبق نتایج حاصل، ۳۳ درصد از عوارض مردم‌گستر دارای نام، نام صحیح، ۴۴ درصد از آن‌ها نام تقریباً صحیح و ۲۳ درصد باقیمانده نام نادرست دارند و دقت توصیفی کل داده‌های مردم‌گستر برابر ۷۷ درصد می‌باشد. البته این نتایج با مقایسه با داده‌های مرجع و با این فرض که نام این داده‌ها صحیح هستند به دست آمده است. در حالی که بررسی‌های صورت گرفته نشان داد این فرض همیشه درست نیست. مثلاً نام بعضی از عوارض مرجع "بدون نام" ثبت شده است در حالیکه عارضه مردم‌گستر متناظر با آن‌ها نام دارد. بنابراین نمی‌توان با قاطعیت اعلام کرد که نام ۲۳ درصد از عوارض مردم‌گستر نادرست است. برای اطمینان از صحت نام این دسته از عوارض می‌بایست آن‌ها را تک تک مورد بررسی قرار داد. همچنین بر اساس نتایج به دست آمده مشخص شد که تمامیت توصیفی مجموعه مردم‌گستر برابر ۴۷ درصد است. بدین معنا که ۴۷ درصد از عوارض مردم‌گستر نام دارند (قلم توصیفی نام برای آنها مقدار دارد).

## ۶- پیشنهاد برای تحقیقات آینده

با استفاده از نتایج حاصل از این تحقیق و با بررسی دقت اطلاعات توصیفی در قالب آنالیزهای مکانی (همچون آنالیز مسیریابی یا راهبری) می‌توان به مطالعه تاثیر دقت توصیفی بر آنالیزهای مختلف مکانی پرداخت. اطلاعات توصیفی نقش انکارناپذیری در اکثر آنالیزهای مکانی دارند و با توجه به استفاده روزافزون از داده‌های مردم‌گستر در سرویس های مکانی، نحوه تاثیر افزایش یا کاهش دقت توصیفی در نتایج نهایی آنالیزها از اهمیت بسزایی برخوردار است.

پیشنهاد دیگر تلاش برای ارائه روشهایی است که به کمک آنها بتوان دقت اطلاعات مردم گستر را بدون نیاز به مقایسه با اطلاعات مرجع محاسبه کرد. چنین روش هایی به ویژه در مناطقی که اطلاعات مرجع برای آن وجود ندارد بسیار مفید و ضروری می باشند.

# مراجع


[1] Bruns, A., (2008). The future is User-Led: The path towards widespread produsage. Fibreculture Journal [Online]. Issue 11.

[2] Goodchild, M.F. (2007). “Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0.” International Journal of Spatial Data Infrastructures Research, 2:24–32

[3] Devillers, R., Stein, A., Bédard, Y., Chrisman, N., Fisher, P., & Shi, W. (2010). Thirty years of research on spatial data quality: achievements, failures, and opportunities. Transactions in GIS, 14(4), 387-400.

[4] Grira, J., Bédard, Y., & Roche, S. (2010). Spatial data uncertainty in the VGI world: Going from consumer to producer. Geomatica, 64(1), 61-72.

[5] Vahedi, B., Alesheikh, A. A., and Honarparvar, S. (2014). Quantitative Assessment of Pragmatic Quality of Volunteered Geographic Information Using Fuzzy Linguistic Quantifiers and OWA Operator. Journal of Geomatics Science and Technology (JGST); 3 (4) :65-76

[6] ISO (International Standardisation Organisation), 2002. ISO 19113:2002 Geographic information — Quality principles.

[7] Devillers, R.; Bédard, Y.; Jeansoulin, R. Moulin, B., 2007. Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data. Transactions in GIS, 21(3):261-282.

[8] Chilton, S. (2009). Crowdsourcing is radically changing the geodata landscape: Case study of OpenStreetMap. In Proceedings of the Twenty-fourth International Cartography Conference, Santiago, Chile

[9] Goodchild, M. F. and Li L. (2012). Assuring the quality of volunteered geographic information. Spatial Statistics 1:110–120

[10] Haklay, M., (2010). How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England, Environment and Planning, 37(4):682-703.

[11] Kounadi, O., (2009). Assessing the quality of OpenStreetMap data. MSc thesis, University College London, UK.

[12] Girres, J-F. and Touya, G., (2010). Quality Assessment of the French OpenStreetMap Dataset. Transactions in GIS, 14(4):435-459.

[13] Ludwig, I., Voss, A. and Krause-Traudes, M., (2010). How Good is OSM? - Method and Results for Germany. In Sixth International Conference on Geographic Information Science 2010, Zurich, Switzerland 14-17 Sep 2010

[14] Koukoletsos, T., Haklay, M., and Ellul, C. (2012). “Assessing data completeness of VGI through an automated matching procedure for linear data”. Transactions in GIS, 16(4), 477-498.

[15] Forghani M., Delavar M., (2014). A Quality Study of the OpenStreetMap Dataset for Tehran. ISPRS International Journal of Geo-Information, 3: 750-763

[16] Mohammadi, N., & Malek, M. (2014). VGI and Reference Data Correspondence Based on Location-Orientation Rotary Descriptor and Segment Matching. Transactions in GIS.

[17] de Smith, M.J., Goodchild, M.F. and Longley, P.A., (2009). Geospatial Analysis - a comprehensive guide: Directional analysis of linear datasets. 3rd edition.

[18] Vahedi, B. (2015). Automatic assessment and presentation of completeness, positional accuracy, and attribute accuracy of linear features in VGI. Master’s thesis, K.N.Toosi University of technology, Tehran, Iran.

[19] Van Oort, P.V. (2006). “Spatial data quality: from description to application”. PhD thesis, Netherlands Geodetic Commission, Delft, The Netherlands

[20] Servigne, S., Lesage, N., & Libourel, T. (2006). Quality components, standards, and metadata. Fundamentals of spatial data quality, 179-210.

[21] Koukoletsos, T. (2012). A Framework for Quality Evaluation of VGI linear datasets. Doctoral dissertation, UCL (University College London).

[22] Levenshtein, Vladimir I, (1966). "Binary codes capable of correcting deletions, insertions, and reversals". Soviet Physics Doklady 10 (8): 707–710.

[23] Ramm, F.; Topf, J. and Chilton, S., (2011). OpenStreetMap Using and Enhancing the Free Map of the World. 3rd ed. Cambridge: UIT Cambridge Ltd.

[24] Li, Z., Zhu, C., & Gold, C., (2010). Digital terrain modeling: principles and methodology. CRC press.

[25] Wikipedia, (2014). http://en.wikipedia.org/wiki/Tehran