

Technische Universität Dortmund

Fakultät Statistik

Bachelorarbeit

**Auswertung von ortsgebundenen Daten auf Straßenebene am Beispiel von
Verkehrsunfällen in London**

eingereicht von: Jakob Richter <jakob.richter@udo.edu>

eingereicht am: 15. Dezember 2012

unter Betreuung: Prof. Dr. Claus Weihs

Inhaltsverzeichnis

1 Einleitung	3
2 Problemstellung	4
2.1 Unfalldatensatz	4
2.2 Problemstellung	5
2.3 Motivation und Problemformulierung	6
2.3.1 Unbekannte Störvariablen	8
2.3.2 Problemformulierung	9
2.4 Statistische Zielsetzung	10
2.5 Zielsetzung zur Visualisierung	11
3 Methoden	11
3.1 Programmietechnische Methoden	11
3.1.1 OSM-Daten in R verfügbar machen	12
3.1.2 Aufbau der Straßendaten	13
3.1.3 Zuordnung der Unfallpunkte zu den Straßen	14
3.1.4 Berechnung der Distanz von Punkten auf den Straßen zu den Unfällen	14
3.1.5 Transformation der Distanzen zur Wahl der inneren und äußeren Umgebung	17
3.1.6 Anmerkungen zur Effizienz	18
3.2 Statistische Methoden	19
3.2.1 Risikodifferenz	19
3.2.2 Binomialtest	20
3.2.3 Exakter Test nach Fisher	20
4 Durchführung	21
4.1 Abweichungen in der Anzahl der in Unfällen verletzten Passanten . . .	22
4.2 Unterschiede der Unfallzahlen in Abhängigkeit der Altersgruppe	23
5 Auswertung und Problematiken	25
6 Zusammenfassung	29
Literatur	33
A Anhang	35
A.1 R-Code	35

1 Einleitung

Die Digitalisierung der Gesellschaft führt zu einer Fülle an Daten. GPS-Empfänger und Smartphones mit verschiedenen Ortungstechniken sind billig und weit verbreitet. Dadurch entsteht eine Vielzahl an ortsbezogenen Daten. Mit solchen Daten befasst sich die *Geospatial Analysis*. Gleichzeitig schreitet die Digitalisierung auch in staatlichen Einrichtungen voran. Unter dem Schlagwort *Open Data* wird immer häufiger gefordert, dass Behörden ihre von Steuergeldern bezahlten Tätigkeiten und insbesondere deren Resultate der Öffentlichkeit verfügbar machen. In den USA ist dies beispielsweise mit dem seit spätestens 1974 etablierten Freedom of Information Act (FOIA) möglich. So ist es dort dank des FOIA jedem Bürger möglich solche Informationen von Behörden einzufordern, welche keinen Geheimhaltungsbedingungen unterliegen. Zur Umsetzung stehen den Bürgern administrative und rechtliche Schritte zur Verfügung. In Deutschland gibt es ein vergleichbares Informationsfreiheitsgesetz (IFG), welches 2006 in Kraft trat; auf Länderebene in Nordrhein-Westfalen bereits schon 2002. Solche Gesetze sichern jedoch lediglich die Möglichkeit auf einen Zugriff. Aktiv sind die Behörden dem IFG nach jedoch nicht dazu verpflichtet, ihre Daten zu veröffentlichen. Dies hat zur Folge, dass sich nur schwer freie Datensätze von allgemeinen Interesse finden lassen. Ungleich mehr und teils sehr umfangreiche Datensätze stellt hingegen die britische Regierung unter ihrem eigenen Open Data Portal <http://data.gov.uk> zur Verfügung.

In dieser Arbeit soll anhand eines Beispiels aufgezeigt werden, wie sich eben solche umfangreichen Daten nutzen und mit anderen Datenquellen verknüpfen lassen um Fragestellungen von allgemeinem Interesse statistisch zu beantworten.

Wo liegen im Londoner Verkehr die Unfallschwerpunkte für Fußgänger?

Dieser Frage nachzugehen ist dank mehrerer durch das Department of Transportation veröffentlichten Datensätze möglich. Diese Daten umfassen Informationen zu allen von der Polizei erfassten Verkehrsunfällen in Großbritannien in den Jahren 2005 bis 2011 und sollen die Grundlage für diese Arbeit bilden, denn insbesondere die darin enthaltene Vielzahl an Variablen bietet einige Untersuchungsmöglichkeiten. So enthalten die Datensätze für jeden Unfall unter anderem Informationen zu der Anzahl der involvierten Fahrzeuge, sowie Anzahl der Unfallopfer, Zeitangaben, Angaben zum Straßentyp und Wetterkonditionen, zudem für jedes Unfallopfer unter anderem Angaben zum Geschlecht, Alter und zur Schwere der Verletzung. Besonders von Interesse sollen für die folgenden Analysen die Angaben zu den Orten der Unfälle sein. Die Angabe der Längen- und Breitengerade ermöglicht es die Daten auf einer Karte zu visualisieren.

Die notwendigen Kartendaten dazu können von dem offenen Projekt *Open Street Map* bezogen werden..

Es soll aufgezeigt werden, wie sich diese umfangreichen Daten durch einen Algorithmus zu einem statistischen Problem transformieren lassen. Von Interesse ist, welche Straßenabschnitte für Fußgänger Unfallschwerpunkte darstellen. Probleme mit denen sich diese Bachelorarbeit beschäftigt sind:

- Wie lässt sich die Anzahl der Unfallopfer visualisieren, sodass daraus ersichtlich wird, welcher Straßenabschnitt einen Unfallschwerpunkt darstellt?
- Wann ist ein Straßenabschnitt als Unfallschwerpunkt einzustufen?
- Wie können Unfallschwerpunkte für bestimmte Personengruppen ermittelt werden?

2 Problemstellung

Zur Beantwortung der in der Einleitung genannten Fragen sei zunächst eine genauere Beschreibung der Daten gegeben. Durch eine einfache Visualisierung dieser Daten wird die Notwendigkeit des im weiteren Verlauf dargestellten Algorithmus und die statistische Auswertung der so erzeugten Daten motiviert.

2.1 Unfalldatensatz

Der Datensatz der durch die Polizei erfassten Unfälle in Großbritannien von 2005 bis 2011 <http://data.gov.uk/dataset/road-accidents-safety-data> besteht für jeweils jedes Jahr aus drei separaten Tabellen. Die Tabelle der Unfälle beinhaltet alle Informationen, die den gesamten Unfall betreffen. Jeder Unfall kommt in ihr genau einmal vor. Daneben enthält die Tabelle der Unfallopfer alle Informationen zu jeder in einen Unfall verwickelten Person. Unfallopfer können mehrmals vorkommen, wenn sie in verschiedene Unfälle verwickelt waren; Sie besitzen keinen eindeutigen Schlüssel. Des Weiteren gibt es eine Tabelle mit den Informationen zu den in den Unfall involvierten Autos sowie Fahrerspezifische Informationen. Diese Daten finden keinen Einzug in die folgende Untersuchung. Die Beobachtungen aus allen drei Tabellen werden durch eine eindeutige Unfall-ID miteinander verknüpft.

Die Daten aus den sechs Jahren werden zusammengefasst und ein zeitlicher Trend wird zu Gunsten größerer Fallzahlen und einer damit einhergehenden höheren Güte außer acht gelassen. Die in den Datensätzen enthaltenen und für die weitere Untersuchung genutzten Variablen finden sich in Tabelle 1.

Tabelle 1: Übersicht, der aus dem Datensatz verwendeten Variablen.

Originalname <i>Übersetzung</i>	Beschreibung	Skalierung
Accident Index <i>Unfallindex</i>	Ein eindeutiger Identifikationsschlüssel für jeden Unfall aus dem Datensatz.	nominal
Location Easting OSGR Location Northing OSGR	Östliche und Nördliche Entfernung vom Nullmeridian im <i>National Grid</i> (OSGB36). Diskretisiert in 10 Einheiten, welche relativ genau 10 Metern entsprechen.	kardinal
Junction Detail <i>Kreuzungsdetails</i>	Angabe an welchem Typ von Kreuzung (z.B. Kreisverkehr, Ausfahrt etc.) der Unfall stattfand oder ob keine Kreuzung in 20 Metern Entfernung ist.	nominal
Age Band of Casualty <i>Alterskategorie</i>	Klassiertes Alter des Unfallopfers in 5 Jahres-Schritten.	kardinal
Casualty Class <i>Unfallopfertklasse</i>	Gibt an, ob das Unfallopfer Fahrer, Pasagier oder Fußgänger ist.	nominal

2.2 Problemstellung

Das Open Street Map (OSM) Projekt ist ähnlich der Wikipedia eine offene Internetplatform. Im Gegensatz zur Aggregation von enzyklopädischem Wissen dient die Open Street Map jedoch zur Sammlung von geografischen Informationen. Bei der Open Street Map handelt es sich also um eine unter freier Lizenz (<http://www.openstreetmap.org/copyright>) stehende benutzergenerierte Weltkarte. Dank der Open Data Commons Open Database License (ODbL) und der OSM-API können die Daten einfach abgerufen und genutzt werden, solange auf die Lizenz und Quelle verwiesen wird und eine Ableitung des Datenbankinhalts unter selber Lizenz veröffentlicht wird. Die in der OSM-Datenbank enthaltenen Daten sind sehr umfangreich. Neben Straßen und deren Typ können unter anderem auch Gebäude, Geschäfte, öffentliche Einrichtungen, Freizeitanlagen und die Art Landnutzung erfasst werden. Von dieser Vielzahl an Daten werden für die weitere Untersuchung in dieser Arbeit nur die für Autos befahrbaren Straßen gewählt. Deren Datengüte ist für den gewählten Kartenausschnitt als gut zu bewerten. Es kann vorkommen, dass manche Straßen unter dem falschen Typ eingetragen sind. Durch die Vielzahl an Typen und deren uneindeutige Benennung tragen Nutzer manche Wege falsch ein. Des weiteren, kann es durch Unachtsamkeit der Benutzer vorkommen, dass manche Straßen nicht miteinander verknüpft sind, obwohl eine Verbindung zwischen beiden besteht. Dies führt dazu, dass ein Algorithmus solche

Übergänge nicht erkennen kann.

2.3 Motivation und Problemformulierung

Eine einfache Visualisierung der Unfälle macht schnell die Unzulänglichkeiten dieser Methoden klar. Werden die Unfallopfer durch Punkte auf der Karte repräsentiert, so ergibt sich ein unübersichtliches Bild (s. Abb. 1), indem sich Punkte überlagern, sodass das Bild keinen deutlichen Informationsgewinn bieten kann.

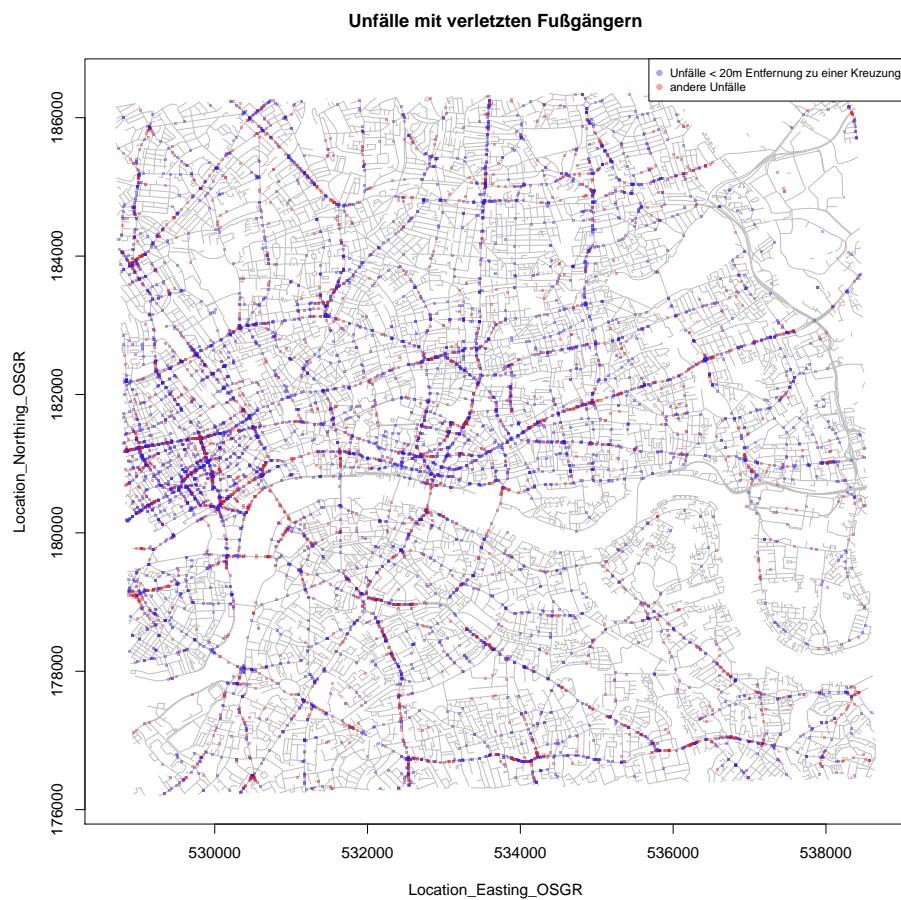


Abbildung 1: Dieser Kartenausschnitt Londons zeigt alle erfassten Unfälle mit involvierten Fußgängern aus den Jahren 2005 bis 2011. Bereits eingezeichnet sind die Straßen aus der *Open Street Map*

Eine weitere Möglichkeit besteht darin, den Ausschnitt in Rechtecke zu unterteilen, die Anzahl der Unfallopfer in den Rechtecken zu zählen und diese entsprechend der Häufigkeiten zu färben. In Abbildung 2 ist leicht zu sehen, dass eine feine Analyse der Straßen so nicht mehr möglich ist. Ein Rechteck mit größerer Straßendichte weiß erwartungsgemäß mehr Unfälle auf. Würde, um dem entgegenzuwirken, das Gitter feiner eingeteilt, so ergäben sich keine erkennbaren Unterschiede mehr. Weitere Probleme sind,

dass ein Gitter ähnlich einer Klasseneinteilung eine Glättung der Daten bedeuten kann, die bestimmte Spitzen verschwinden lässt. Außerdem dominieren die Bereiche mit den meisten Unfällen im Kartenausschnitt die Grafik, sodass lokale Maxima in Bereichen mit weniger Unfällen kaum sichtbar sind.

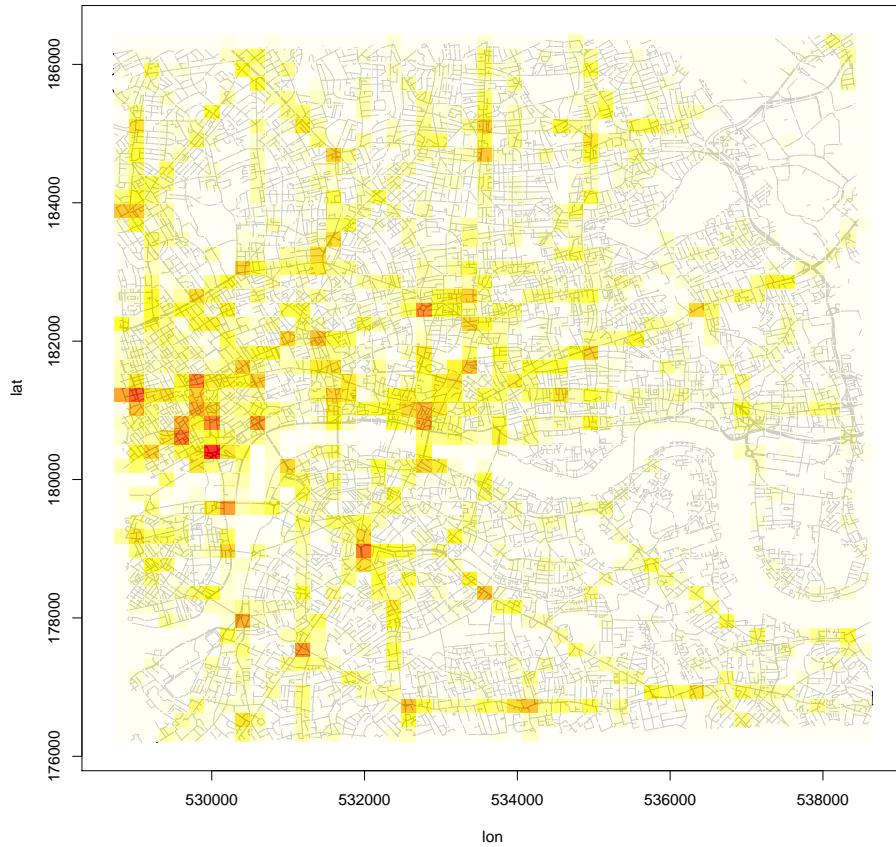


Abbildung 2: Der selbe Kartenausschnitt mit den selben Unfalldaten wie in Abb. 1 zeigt nun die Unfallhäufigkeit in über die Karte gelegten Quadraten an. Ein dunkles Rot indiziert hohe Unfallzahlen.

Die Bindung von Verkehrsunfällen an Straßen lässt es sinnvoll erscheinen, die Idee des Histogramms mit der Straßendarstellung einer Karte zu kombinieren. Straßen mit vielen Unfällen werden also entsprechend gefärbt. Das später noch genauer vorgestellte Verfahren erfasst die Unfälle und ordnet sie den Straßen zu. Gegensätzlich zum Histogramm und zur Abbildung 2 lässt sich jedoch keine feste Klassenbreite wählen, was hier einer festen zu untersuchenden Straßenlänge entsprechen würde, da Kreuzungen die Straßen sehr unregelmäßig und in kurzen Abschnitten zerteilen. Es wird ein Umweg gegangen: Die Färbung jedes Straßensegmentes gibt die Anzahl der Unfälle an, welche innerhalb einer festgelegten Laufweite zu erreichen sind. Es ist also nicht die Luftliniendistanz, sondern die auf der Straße zurückgelegte. Die so Erzeugte Abbildung 3 zeigt die absolute Anzahl in Verkehrsunfällen verletzten Fußgänger innerhalb einer Umgebung von 500

Metern Laufweite. Es wird deutlich, dass es großflächige Gegenden mit hohen Unfallzahlen gibt, innerhalb derer es nicht möglich ist, spezielle Unfallschwerpunkte zu erkennen. Es wird also nötig sein solche Gebiete zu unterscheiden.

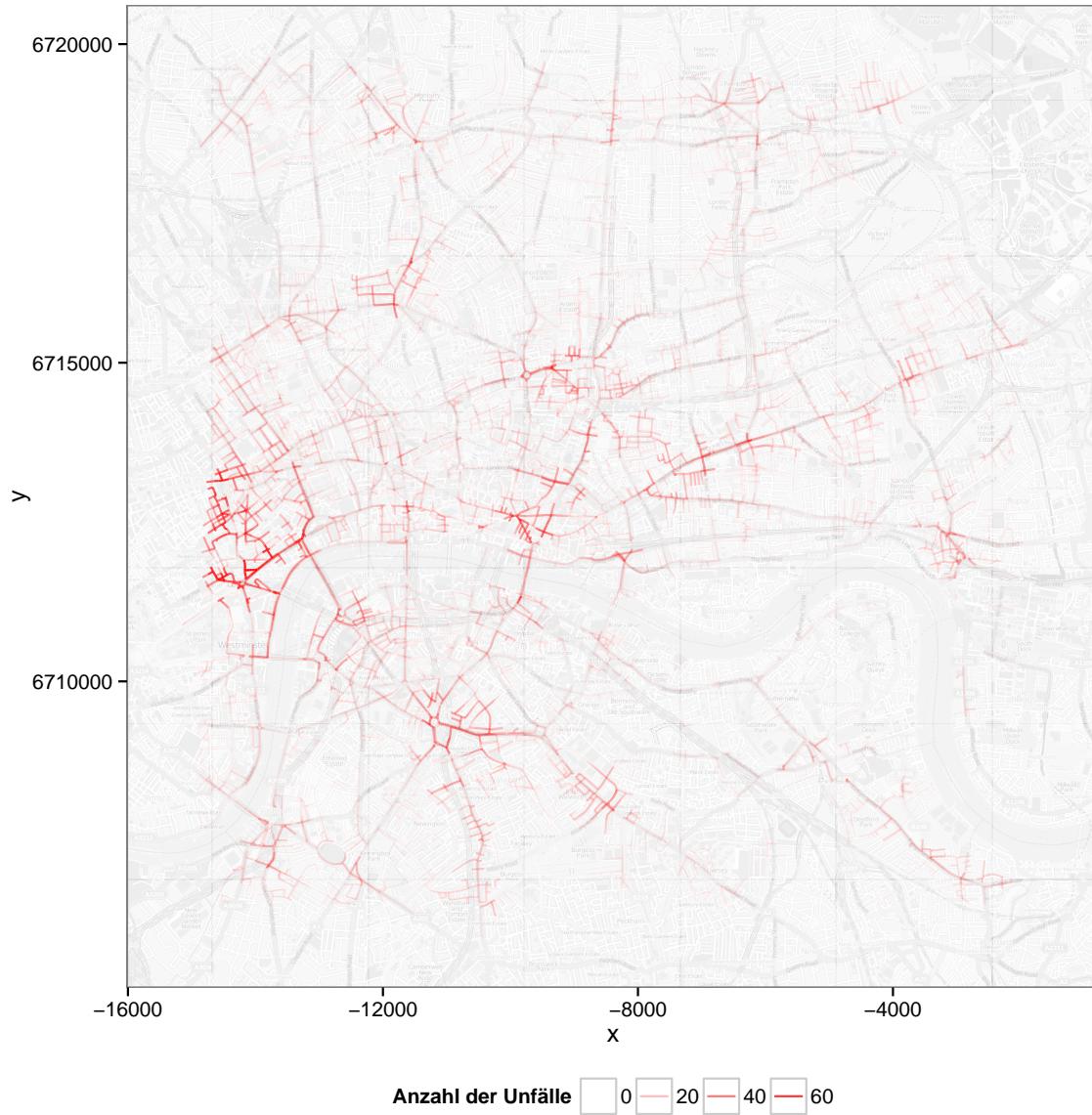


Abbildung 3: Die Intensität des Rots an jeder Straßenstelle gibt die Anzahl der in Verkehrsunfällen involvierten Passanten an, welche sich in einer Laufweite von bis zu 500 Metern befinden.

2.3.1 Unbekannte Störvariablen

Das Problem der unbekannten störvariablen steht in Verbindung mit der Fragestellung dieser Arbeit, was als Unfallschwerpunkt angesehen werden soll. Denn auch bei obiger Methode (vgl. Abb. 3) die Anzahl der Unfälle darzustellen, bleiben einige Variablen unberücksichtigt und erlauben damit kein Rückschluss auf die Fragestellung, ob es sich

bei dem betrachteten Punkt um einen Unfallschwerpunkt handelt. So ist zu vermuten, dass die Anzahl der durch Verkehrsunfälle verletzten Passanten auf einer Straße stark von der Anzahl der Autos abhängt, die im Durchschnitt pro Tag auf dieser fahren und ebenfalls stark von der Anzahl der Fußgänger abhängt, die diese Straße überqueren. Diese Daten liegen nicht vor, gehören damit zu den unbekannten Störvariablen. Das nachträgliche Erheben dieser Variablen für alle Straßen ist sehr aufwendig, darum soll die folgende statistische Auswertung einen Weg aufzeigen, den Einfluss dieser Störvariablen zu verringern:

Für eine genügend kleine betrachtete Umgebung wird davon ausgegangen, dass die Ausprägungen der Störvariablen ‘Fußgängerfrequentierung’ und ‘Kraftfahrzeugfrequentierung’ innerhalb der Umgebung und im direkten Umfeld nahezu gleich sind. Ebenso kann für hier nicht genannte Störvariablen argumentiert werden.

Das Vorgehen stets nur eine kleine Umgebung zu betrachten verhindert gleichzeitig, dass die größten Unfallhäufigkeiten an bestimmten Stellen das Bild dominieren, sodass das Ergebnis unabhängig von der Größe des gewählten Kartenausschnitts ist.

2.3.2 Problemformulierung

Für einen Punkt wird eine innere Umgebung und eine diese umgebende äußere Umgebung zu finden sein, sodass diese verglichen werden können. Diese Umgebungen sollten so wählbar sein, dass beide die gleiche Straßenstrecke überdecken.

Ein Straßenabschnitt wird als *Unfallschwerpunkt* definiert, wenn er und seine nähere Umgebung signifikant mehr Unfälle aufweist, als eine die innere Umgebung umgebende äußeren Umgebung und beide Umgebungen die gleiche Straßenstrecke überdecken.

Für eine Personengruppe wird ein Straßenabschnitt als Unfallschwerpunkt definiert, wenn in der inneren Umgebung der Anteil der Unfälle mit Personen aus einer Personengruppe an allen Unfällen in dieser Umgebung signifikant höher ist als der selbe Anteil in der äußeren Umgebung.

Ein Algorithmus wird die Angaben über Längen- und Breitengerade eines jeden Unfalls auf die Position auf den Straßen abbilden. Für einen Punkt auf der Straße werden die Unfälle ermittelt, welche innerhalb einer inneren und äußeren Umgebung um diesen Punkt liegen. Für einen Vergleich der absoluten Unfallzahlen müssen beide Umgebungen die gleiche Größe haben. Dies soll so geschehen, wie es Abbildung 4 veranschaulicht.

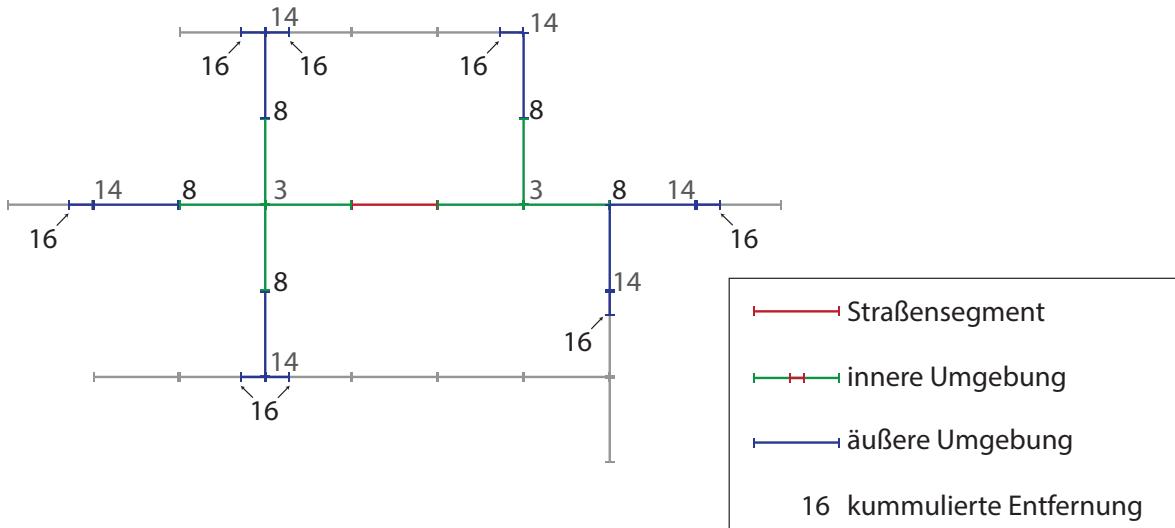


Abbildung 4: Schema zur Veranschaulichung des hier genutzten Konzepts der Umgebung. Die Zahlen geben an, wie groß eine Umgebung bis zu diesem Punkt wäre. Der Wert 16 am Rand der äußeren Umgebung zeigt also die Größe der gesamten Umgebung bis dorthin an. Durch die Wahl der inneren Umgebung bis zu den Punkten, an denen die 8 erreicht ist, überdecken äußere und innere Umgebung die selbe Streckenlänge.

2.4 Statistische Zielsetzung

Mit den gleich großen Umgebungen lässt sich für zwei Umgebungen Hilfe des *Binomialtests* untersuchen, ob die Differenz der Unfallopfer zwischen den Umgebungen signifikant ist. Da innere und äußere Umgebung ein Straßengebiet der gleicher Straßenlänge abdecken, erwarten wir unter der Nullhypothese, dass die innere Umgebung keinen Unfallschwerpunkt darstellt, dass die Wahrscheinlichkeit für einen Unfall in der inneren Umgebung zu liegen kleiner gleich der Wahrscheinlichkeit ist in der äußeren Umgebung zu liegen.

Anstatt nur die Anzahl der Unfälle zu untersuchen, lassen sich Tests auch mit bestimmten Personengruppen durchführen, welche durch die Variablen im Unfallopferdatensatz unterscheidbar sind.

Für eine bestimmte Personengruppe wird sowohl die Anzahl der Unfallopfer der spezifischen Personengruppe in den Umgebungen als auch die Anzahl aller anderen Unfallopfer zum Testen herangezogen. Ist ein Punkt als Unfallschwerpunkt für eine spezifische Personengruppe anzusehen ist dies mit der obigen Definition nach **Hartung** mit der positiven stochastischen Abhängigkeit der Merkmale „innere Umgebung“ und „Personengruppe“ äquivalent. Für die Untersuchung in dieser Arbeit sei folgende Personengruppe gewählt: Alle Fußgänger, und Kinder im Alter zwischen 6 und 15. Dies soll insbesondere Unfallschwerpunkte für Schulkinder aufzeigen.

Das Vorliegen einer positiven stochastischen Abhangigkeit kann mit dem einseitigen *Exakten Test nach Fischer* uberpruft werden.

2.5 Zielsetzung zur Visualisierung

Es soll eine Straenkarte entstehen, aus welcher ersichtlich ist wo potentielle Unfallschwerpunkte liegen. Dies wird so umgesetzt, dass die Straen in kleine Straensegmente unterteilt werden. Von dem Mittelpunkt jedes Straensegments ausgehend werden die Berechnungen durchgefuhrt. Die Resultate fur den Mittelpunkt dienen dann zur Farbung des ganzen Segments. Bei kleinen Segmenten kann so davon ausgegangen werden, dass das Ergebnis approximativ fur jeden Punkt auf dem Segment gilt. Fur jedes Segment sollen zwei Werte vorliegen. Die Signifikanz des Testergebnisses entscheidet daruber, ob und wie intensiv das Segment gezeichnet wird. Eine weitere Kennzahl ist Grundlage zur Farbung des Segments.

3 Methoden

In diesem Kapitel sollen die genutzten statistischen Tests genauer erlautert werden, sowie insbesondere die Datenstrukturen und das algorithmische Vorgehen. Zum einen wird dargelegt, wie mit der Programmiersprache *R* (R Core Team 2012) und dem R-Paket *osmar* (Eugster und Schlesinger 2010) sowie den Programmen *osmosis* (*Osmosis - OpenStreetMap Wiki*) und *osmfilter* (*Osmosis - OpenStreetMap Wiki*) die OSM-Daten erst auf wesentliche Informationen reduziert werden und wie diese dann in *R* eingelesen werden. Die Weiterverarbeitung in *R* wird anhand der Beschreibung der verwendeten Algorithmen gegeben. Der dazugehorige Quelltext findet sich im Anhang ???. Einige Grafiken werden mit dem R-Paket *ggplot2* (Wickham 2009) erstellt. Das Einbinden der grafischen OpenStreetMap-Karte in die Grafiken erfolgt mit Hilfe des Pakets *OpenStreetMap* (Fellows und Jan Peter Stotz 2012).

3.1 Programmietechnische Methoden

Im Folgenden soll nun vorgestellt werden, wie die Kartendaten der Open Street Map bezogen und verarbeitet werden konnen, sodass sie in *R* eingelesen werden konnen. Danach wird der Aufbau der Straendaten kurz erlautert und das Vorgehen beschrieben, wie die Unfallpunkte auf das Straennetz abgebildet werden, sowie das Vorgehen statistisch auswertbare Daten fur jeden Straßenabschnitt zu erhalten. In der Erlauterung wird sich an die Datenlage der behandelten Problematik orientiert. An die Stelle

der hier verwendeten *Unfallpunkte* können selbstverständlich auch andere Punkte treten.

3.1.1 OSM-Daten in R verfügbar machen

Schritt für Schritt sei hier zunächst erläutert, wie vorgegangen wurde um an die OSM-Daten zu gelangen. Das R-Paket osmar bietet eine direkte Schnittstelle zur OSM-API an, sodass aus R heraus auf die stets aktuellsten OSM-Datenbankinhalte zugegriffen werden kann. Auf Grund der großen Datenmenge für den gewählten Kartenausschnitt und da lediglich die für Autos freigegebenen Straßen von Interesse sind wurde ein anderer Weg gewählt.

Zuerst wird eine osm-Datei für die Region von Interesse bezogen. Die Quelle für die Daten dieser Arbeit ist *Downloads von OSM-Dateien*.

Mit dem Programm *osmosis* (*Osmosis - OpenStreetMap Wiki*) (benötigt Java Runtime Environment) kann diese Datei nach dem Entpacken weiter auf den spezifischen Kartenausschnitt verkleinert werden. Folgende Kommandozeile generiert so aus der OSM-Datei eine neue, welche nur noch die Punkte enthält, welche in gegebenen Rahmen liegen.

Listing 1: Konsolenbefehl zum Ausführen von osmosis (benötigt JRE)

```
1 osmosis --read-xml enableDateParsing=no file="united_
kingdom.osm" --bounding-box top=51.575 left=-0.275
bottom=51.445 right=0.015 --write-xml file="london_inner
.osm"
```

Danach wird das Programm *osmfilter* (*Osmfilter - OpenStreetMap Wiki*) genutzt um nur bestimmte Elemente aus der OSM-Datei in eine neue zu übernehmen. Es werden die „Highway“- Typen gewählt.

Listing 2: Genutzter Konsolenbefehl zum Beschränken der Straßentypen mit Hilfe von osmfilter.exe

```
1 osmfilter.exe "london_inner.osm" --keep= --keep-ways="
highway=crossing =living_street =mini_roundabout =
motorway =motorway_junction =motorway_link =primary =
primary_link =residential =road =secondary =secondary_
link =tertiary =tertiary_link =track =trunk =trunk_link
=turning_circle =unclassified =unsurfaced" -o="london_
inner_streets_small.osm"
```

Die nun entstandene Datei ist klein genug um sie ohne Performanceprobleme unter Verwendung des Pakets *osmar* mit R einzulesen und ein Rdata-File für spätere Verwendung zu speichern.

Listing 3: Einlesen der erstellten osm-Datei in R mit Hilfe des Paketes *osmar*.

```

1 library(osmar)
2 raw <- readLines("london_inner_streets_small.osm")
3 gbxml <- xmlParse(raw)
4 umgebung <- as_osmar(gbxml)
5 save(umgebung ,file="umgebung.Rdata")

```

In den nun erzeugten *osmar*-Objekt *umgebung* liegen die Angaben über die Lage der Punkte, aus denen die Straßen bestehen im weit verbreiteten WGS 84 Datumsformat vor. Der Abstand zwischen zwei Längengeraden unterscheidet sich jedoch in Europa stark von dem zweier Breitengerade. Für vereinfachte Rechnungen werden darum die Punkte in R mit dem Paket *rgdal* Keitt et al. 2012 auf das Koordinatensystem des National Grid (OSGB36) projiziert, welches im Raum Großbritanniens so geartet ist, dass eine Einheit nördlicher Richtung ebenso wie eine Einheit östlicher Richtung einem Meter entspricht, was weitere Rechnungen erleichtert.

3.1.2 Aufbau der Straßendaten

Der genaue Aufbau des *osmar*-Objekts *umgebung* kann Eugster und Schlesinger (2010) entnommen werden. Durch oben genannte Schritte enthält das *osmar*-Objekt *umgebung* nur Straßendaten. Für den später folgenden Algorithmus werden diese Daten so benötigt, dass jede Straßenkreuzung und jedes Straßenende einen Knoten im Sinne der Graphentheorie bildet. Die in den OSM-Daten vorliegenden *ways* repräsentieren Straßen und Straßenteile und sind Linien die durch mehrere Punkte (*nodes*) führen, welche auch Knoten sein können. Folglich ist eine Zerstücklung der *ways* in Kanten notwendig. (Anmerkung: Es kann in den OSM-Daten vorkommen, dass ein node genau zu zwei *ways* gehört, welche in ihm Enden. Dieser Knoten ist demnach keine Straßenkreuzung. Dies wird im Folgenden vernachlässigt.) Abweichend von der Graphentheorie sind die Kanten in dieser Arbeit Linien von Knoten zu Knoten, welche auch aus Zwischenpunkten (den *nodes*) bestehen, da der genaue Straßenverlauf von Kreuzung zu Kreuzung für die geographische Auswertung wichtig ist. Im weiteren Verlauf dieser Arbeit werden eben diese Kanten als Straßen bezeichnet.

3.1.3 Zuordnung der Unfallpunkte zu den Straßen

Nachdem die eigene Datenstruktur aufgebaut ist, kann nun begonnen werden, für alle Straßen die Unfälle zu suchen, die auf ihr liegen. Dazu wird über alle Straßen iteriert. Jede dieser einzelnen Straßen besteht aus mehreren Strecken (siehe Abb. 5) über welche wiederum iteriert wird. Für jede Strecke $S_i = ((x_i, y_i), (x_{i+1}, y_{i+1}))$ wird dann die Neigung α_i berechnet. Mit dem so errechneten Winkel α_i kann nun die Transformationsmatrix M_i berechnet werden, welche die Strecke so rotiert, dass sie horizontal liegt.

$$\begin{aligned} \text{Steigung der Strecke } i: \quad \alpha_i &= \tan^{-1} \left(\frac{y_i - y_{i+1}}{x_i - x_{i+1}} \right) \\ \text{Rotationsmatrix:} \quad M_i &= \begin{pmatrix} \cos(\alpha_i) & -\sin(\alpha_i) \\ \sin(\alpha_i) & \cos(\alpha_i) \end{pmatrix} \\ \text{Rotation der Punkte:} \quad (\tilde{x} \ \tilde{y}) &= (x \ y) \cdot M \end{aligned}$$

Die selbe Transformation wird mit den Punkten, welche die Unfälle repräsentieren, vollzogen. Nun sind die Punkte in Straßennähe für die jeweilige Strecke leicht durch die Lage in einem Rechteck zu wählen. $U_i = \{P(x, y) | x > x_1 - h_1 \wedge x < x_2 + h_2 \wedge y < y_l + h \wedge y > y_l - h\}$ mit $h_1 = 0$, wenn $i = 1$ und $h_2 = 0$, wenn $i = n$. Die Position des Unfalls auf der Straße wird für den Punkt $P(x, y)$ definiert als $d := x - x_1$. Sollte ein Punkt durch Überlagerung der Rechtecke auf mehreren Positionen d_1, \dots, d_k der Straße liegen so wird die endgültige Distanz d auf der Straße gemittelt $d = \frac{1}{k} \sum_{i=1}^k d_i$. Die Ergebnisse werden in einer Liste gespeichert, sodass für jede Straße eine Tabelle vorliegt, welche die Unfälle mit ihrer eindeutigen ID und der Entfernung vom Startpunkt der Straße enthält. Es kann vorkommen, dass zwei Straßen so dicht nebeneinander liegen, dass Unfälle mehreren Straßen zugewiesen werden. Dieser Effekt hängt von dem gewählten Suchabstand h ab und wird zugunsten der Effizienz vernachlässigt. Außerdem wird er durch den im weiteren vorgestellten Algorithmus Marginalisiert, da dieser bei mehrfachen Vorkommens eines Unfalls stets den mit der kürzesten Distanz zum Ausgangspunkt wählt.

3.1.4 Berechnung der Distanz von Punkten auf den Straßen zu den Unfällen

Nun, da für alle Straßen die Unfälle mit ihrer Entfernung vom Straßenanfang bekannt sind, soll der Algorithmus vorgestellt werden, welcher für jede Kreuzung die Unfälle

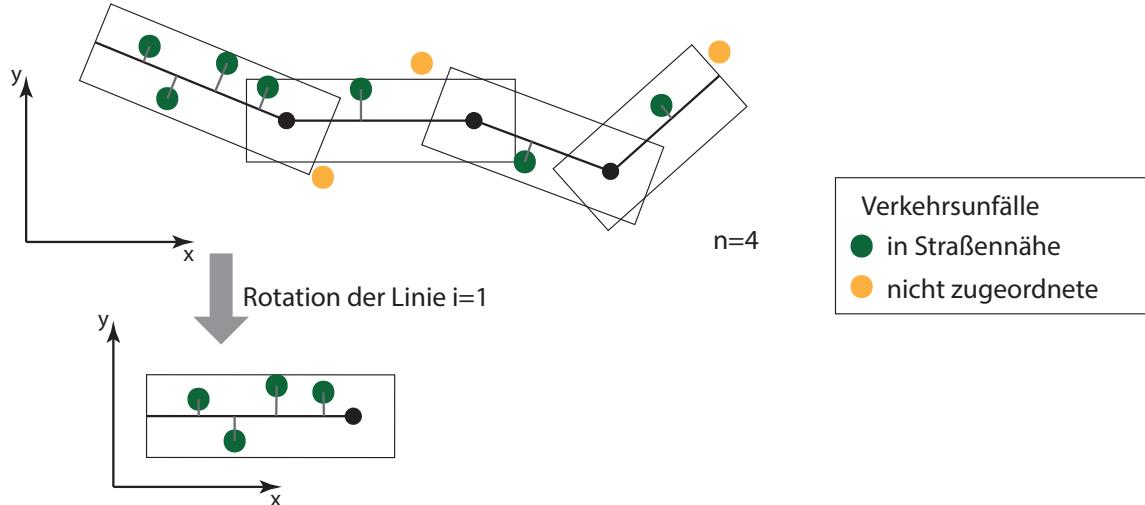


Abbildung 5: Um herauszufinden, welche Unfälle in Straßennähe liegen, werden die Straßenabschnitte so gedreht, dass die Auswahl über eine Selektion der x und y Werte geschehen kann.

innerhalb eines vorgegebenen Suchradius findet. In der folgenden Erläuterung werden im Code umgesetzte Effizienzverbesserungen vorerst nicht geschildert. Es wird über alle Kreuzungen (Knoten) iteriert, sodass die Suche von jeder Kreuzung aus anfängt.

Für den Algorithmus zur Suche der kürzesten Strecke bis zu einem Unfall auf einer Graphenähnlichen Struktur $G(V, E)$ werden folgende Variablen genutzt.

- $V = v_1, \dots, v_r$ $r \in \mathbb{N}$, Menge aller Knoten des Graphen.
- $E = e_1, \dots, e_s$ $s \in \mathbb{N}$, Menge aller ungerichteten Kanten des Graphen.
- l , bisher gelaufene Strecke.
- l^{max} , maximal zu laufende Strecke.
- $S(v_l)$ $l = 1, \dots, r$ ist definiert als bisher kürzeste Distanz von Startknoten zum Knoten v_l . Wurde der Knoten v_l noch nie zuvor besucht, so beträgt $S(v_l) = Inf$.
- d_i^{total} , Laufweite vom Unfall zum Startknoten v_1 .
- $L(e_k)$, Länge der Straße, welche durch die Kante e_k repräsentiert wird.

Für die Suche vom Knoten v aus, arbeitet der Algorithmus wie folgt:

1. Stelle fest ob der Knoten schon besucht wurde. Ist $S(v) < l$ stoppe die Suche, da der Knoten bereits durch einen kürzeren Weg erreicht wurde. Ansonsten fahre fort.
2. Setze $S(v) := l$.

3. Suche vom Knoten v abgehende Kanten e_1, \dots, e_k mit $e_i = v_1, w_i$.
4. Rufe für alle Kanten e_1, \dots, e_k die Methode zum Ablaufen einer Kante auf. Der Wert der bisher gelaufenen Strecke beträgt l .

Die Methode zum Ablaufen einer Kante $e = v, w$ vom Knoten v besteht aus folgenden Schritten:

1. Die bisher zum Knoten v gelaufene Strecke l ist bekannt.
2. Für die Kante e sind die Entfernungen der Unfälle d_i entweder von Knoten v oder w bekannt, sowie ihre Länge $L(e_k)$. Sind die Entfernungen vom Knoten v bekannt, so entspricht die Laufweite vom Startknoten zum Unfall $d_i^{total} = l + d_i$. Im ungekehrten Fall $d_i^{total} = l + (s - d_i)$.
3. Ist $l + s >= l^{max}$, so ist die Suche an dieser Stelle beendet.
4. Ist $l + s < l^{max}$, so starte Suche vom Knoten w aus und setze $l := l + s$.

Der Algorithmus startet mit dem Knoten v_1 , welcher die Kreuzung repräsentiert, von der aus die Entfernung gemessen werden soll, bei bisheriger Laufweite $l = 0$. Nun sind von allen Kreuzungen die Entfernungen zu den Unfällen innerhalb der maximalen Laufweite l^{max} bekannt. Für die Unfälle, für die am Ende mehrere Wege gefunden wurden, wird sich für den kürzesten Weg entschieden und die anderen Ergebnisse fallen weg.

Es ist abgesichert, dass

- stets der kürzeste Weg zum Unfall gefunden wird,
- kein Unfall doppelt vorkommt,
- der Algorithmus nicht im Kreis geht und somit auch, dass
- der Algorithmus bei Erreichen der maximalen Laufweite terminiert.

Nun ist noch der Abstand zu bestimmten Punkten auf der Straße von Interesse. Die Feinheit h , alle wie viele Meter die Distanzen von einem Punkt gemessen werden soll, kann hierbei angepasst werden. Die Zwischenpunkte auf der Straße werden so gelegt, dass sie der vorgegebene Feinheit h am nächsten kommen. Für alle Straßen im gewählten Kartenausschnitt werden die Zwischenpunkte berechnet. Die Anzahl z der Zwischenpunkte auf einer Straße der Länge s beträgt: $z =: \lfloor \frac{s}{h} \rfloor$. Die tatsächliche Feinheit beträgt damit $\tilde{h} = \frac{s}{z}$ und kann damit vom gewünschten h um $\frac{h}{2}$ abweichen. Es wird über alle Zwischenpunkte iteriert. Sei für einen Durchgang p ein Zwischenpunkt

auf Kante $e = v, w$ mit der Entfernung s_v zu dem Kantenende v , dann fahre wie folgt vor:

1. Die Distanzen d_i zu Unfällen auf der Straße der Länge $L(e)$ (Der Weg von v nach w), welche durch die Kante e repräsentiert wird berechnen sich mit $d_i = L(e) - s_v$.
2. Die Distanzen zu Unfällen, die durch den Knoten v erreicht werden, berechnen sich aus den bereits bekannten Unfalldistanzen zu diesem Knoten d_{v_i} mit $d_i = d_{v_i} + s_v$.
3. Für die Distanzen zu Unfällen über den Knoten w ergibt sich ähnlich $d_i = d_{w_i} + (s - s_v)$
4. Sollten für ein Unfall mehrere Distanzen vorliegen, wird sich für die kürzeste entschieden.

3.1.5 Transformation der Distanzen zur Wahl der inneren und äußeren Umgebung

Dieses vorgehen sichert bis jetzt vorerst, dass für alle Zwischenpunkte die Distanzen zu den Unfällen innerhalb des Suchradius bekannt sind. Bei einem dichten Straßennetz Liegen innerhalb der maximalen Laufweite l^{max} jedoch mehr Straßen, als in einem Gebiet mit nur wenigen Straßen. Um jedoch sicherstellen zu können, dass die untersuchte Straßenlänge für alle Punkte die selbe ist, damit die Voraussetzungen für die statistischen Auswertungen gegeben sind, muss der Algorithmus erweitert werden. Die Idee besteht darin, dass bei bildlicher Vorstellung, sich der Algorithmus in alle Richtungen konstant ausbreitet und sich an Kreuzungen aufspaltet. Ist die Suche die Distanz l bis zu einem Knoten v mit k abgehenden Kanten gelaufen, läuft sie von dort aus auf $k - 2$ abzweigenden Kanten weiter. $k - 2$, da sie von einer der k abgehenden Kanten zu dem Knoten gelangt ist und eine Kante die Weiterführung der bisherigen Suche ist. Handelt es sich um einen Endknoten hat er $k = 1$ abgehende Kanten und die Suche wird auf -1 zusätzlichen Kanten, sprich einer Kante weniger weitergeführt. Diese Information ist für die parallel noch weiter laufenden Suchen wichtig. Die Suche deckt ab der Laufweite l die $k - 1$ -fache Strecke für die weitere Distanz l_n ab. l_n ist die Distanz bis eine der $k - 1$ Suchen auf einen Knoten mit k_n abgehenden Kanten trifft. Der Multiplikator beträgt ab l_n dann $1 + (k - 2) + (k_n - 2)$. Dieses Prinzip soll auch Abb. 6 veranschaulichen.

Zur Umsetzung werden alle Abzweigungen mit ihrer Distanz zum Startpunkt während des Durchlaufens des Algorithmus gespeichert. Sind für den Knoten alle Unfälle mit ihrer kürzesten Distanz gefunden kann so die kumulierte Distanz d_c berechnet werden. $A(s)$ sei die Anzahl der Abzweigungen, die der Algorithmus nach der Laufweite s

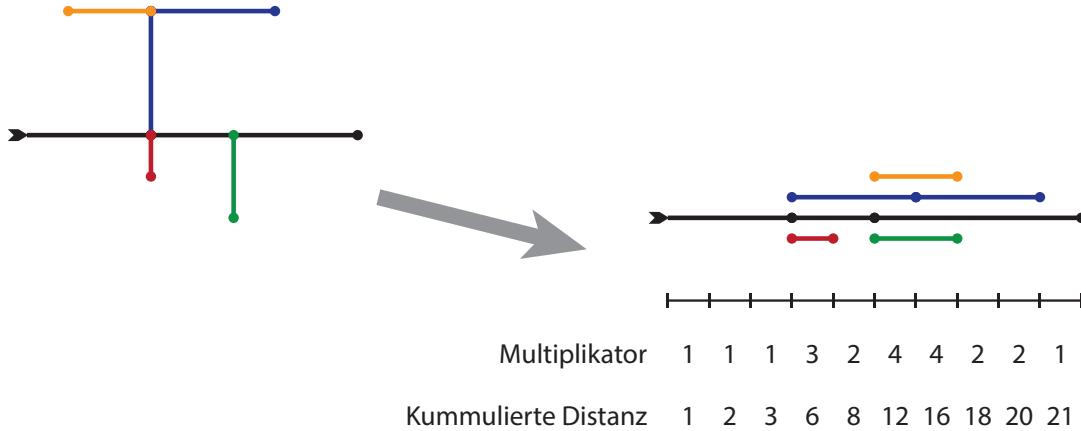


Abbildung 6: Diese Skizze veranschaulicht, wie die kummulierte Distanz von einem Startpunkt (Pfeispitze) aus berechnet wird.

gefunden hat. s_1, \dots, s_k mit $0 = s_1 < s_2 \leq \dots \leq s_k$ seien die Laufweiten, an denen Abzweigungen vorliegen. $A_c(s)$ ist dann die Summe der Abzweigungen bis zu der Laufweite s : $A_c(s) := \sum_{s_i, s_i < s} s_i$. Dann lässt sich für einen Unfall mit einer kürzesten Distanz d und $s_a < d < s_b$, $a, b \in 1, \dots, k$ die kummulierte Distanz d_c wie folgt berechnen: $d_c = A_c(s_1) \cdot s_1 + A_c(s_2) \cdot s_2 + \dots + A_c(s_a) \cdot s_a + A_c(s_b) \cdot d$.

Wird die innere Umgebung so gewählt, dass sie nur Unfälle mit kumulierten Distanzen $d_c \leq \frac{l^{max}}{2}$ enthält und die äußere Umgebung so, dass diese nur Unfälle mit kumulierten Distanzen d_c mit $\frac{l^{max}}{2} < l^{max}$ enthält ist somit sichergestellt, dass beide Umgebungen die selbe Straßenstrecke der Länge $\frac{l^{max}}{2}$ abdecken.

Da die Laufweite l^{max} auch zu den Rändern hin ausgenutzt werden sollte um unverfälschte Ergebnisse zu erhalten, werden nur die Ergebnisse von Knoten mit einem Abstand von l^{max} zum gewählten Rand des Kartenausschnitts berücksichtigt.

3.1.6 Anmerkungen zur Effizienz

Die Effizienz soll in dieser Arbeit nicht von großer Bedeutung sein, dennoch sollen einige Anmerkungen gegeben sein, da kleine Verbesserungen auch in den Code eingeflossen sind. Eine Schwachstelle des Algorithmus ist, dass bei den Iterationen über alle Knoten für jede Knotensuche teilweise aus der gleichen Richtung kommend die gleichen Knoten abgelaufen werden. Um dies zu vermeiden wäre ein hoher Grad an Rekursivität nötig oder eine komplexe iterative Lösung. Der im Code gewählte Ansatz basiert auf der Idee, dass für einen bereits fertig berechneten Knoten das Ergebnis von ihm aus gesehen bereits optimal ist. Wird also in einer späteren Iteration dieser Knoten im Laufe der Suche erreicht, so kann an der Stelle gestoppt werden und die Ergebnisse des bereits

berechneten Knoten werden mit den aktuellen Ergebnissen entsprechend vereint. Gibt es in der Umgebung des Startknotens keinen Unfall in direkter Luftliniendistanz von l^{max} , so kann es auch keinen Unfall in Laufweite von l^{max} geben und die Suche kann beendet werden. Dazu muss die Distanz der Unfälle nicht berechnet werden, sondern es wird über dem Unfalldatensatz eine Selektion ausgeführt, sodass diese nur Unfälle innerhalb eines Rechtecks mit entsprechenden Ausmaßen beinhaltet. Ist die Selektion leer, so ist dies auch das Ergebnis der Suche auf dem Knoten. Der Speicherbedarf des Algorithmus ist relativ hoch. Für einen $100km^2$ großen Bereich um das Zentrum Londons wurden knapp 6 GB Hauptspeicher belegt. Eine Optimierungsmöglichkeit ist dem Autor nicht bekannt. Es ist jedoch denkbar, für größere Ausschnitte die Ergebnisse für kleinere Teilkacheln zu berechnen. Dies ist jedoch sehr umständlich, da die Kacheln mindestens eine Ausdehnung von $2 \cdot l^{max}$ haben und sich auf allen Seiten über eine Strecke von l^{max} Überschneiden sollten.

3.2 Statistische Methoden

Die Vorarbeit des Algorithmus bereitet Möglichkeit zur Auswertung auf bekannte statistische Tests zurückzugreifen. Zum einen soll der Binomial-Test vorgestellt werden sowie der exakte Test nach Fisher. Neben den Tests werden jedoch auch Kennzahlen benötigt welche die Unterschiede zwischen den Umgebungen ausdrücken.

3.2.1 Risikodifferenz

Gegeben sei eine Vierfeldertafel mit den Bezeichnungen wie in Tabelle 2. Es gibt N Beobachtungen und zwei erfasste dichotome Merkmale A und B . Die Risikodifferenz ist dann nach **Hartung** wie folgt definiert:

$$RD := \frac{n_{11}}{n_{1\cdot}} - \frac{n_{21}}{n_{2\cdot}} \quad (1)$$

Ist einer der Zähler (n_{11}, n_{21}) gleich Null oder gleich dem Nenner so wird dem Zähler 0,5 hinzugaddiert bzw. im zweiten Fall 0.5 abgezogen. Diese Kennzahl nimmt Werte zwischen -1 und $+1$ an. Ein Wert nahe der 0 zeigt einen kleinen Unterschied der Risiken zwischen den beiden Gruppen $B = 0$ und $B = 1$. Ein negativer Wert zeigt an, dass in der Gruppe $B = 1$ anteilig weniger Merkmalsträger mit der Ausprägung $A = 1$ vorkommen als in der Gruppe $B = 0$. Bei einem positivem Wert ist analog der Anteil der Merkmalsträger mit $A = 1$ in der Gruppe $B = 1$ größer.

Diese Kennzahl bietet den Vorteil, dass sie Werte innerhalb eines festen Intervalls annimmt, sodass sie als Grundlage für eine Farbskala dienen kann. Außerdem ist es

wichtig, dass sie auch definiert ist, wenn es Zellen in der Vierfeldertafel gibt, in die keine Beobachtung fällt. Dieses Kriterium erfüllen weder das *Relative Risiko* noch das *Odds Ratio*.

Tabelle 2: Vierfeldertafel

		A	1 0	Σ
		B		
		1	$n_{11} \quad n_{12}$	$n_{1\cdot}$
		0	$n_{21} \quad n_{22}$	$n_{2\cdot}$
		Σ	$n_{\cdot 1} \quad n_{\cdot 2}$	N

3.2.2 Binomialtest

Der einseitige Binomialtest nach Rüger 2002[S. 14 ff.] für das Hypothesenpaar $H_0 : p \leq p_0$ gegen $H_1 : p > p_0$ betrachtet die Stichprobe $X = (X_1, \dots, X_n)$ aus n unabhängig identisch bernoulli-verteilten Zufallsvariablen X_i mit $X_i \sim Bern(p)$ und $i = 1, \dots, n$, wobei $p = P(X_i = 1)$ unbekannt. Die Teststatistik $Z = \sum_{i=1}^n X_i$ ist binomialverteilt mit $Z \sim Bin(n, p)$. Der p -Wert für den beobachteten Wert c der Teststatistik beträgt:

$$p(c) = P_{H_0}(Z \geq c) \tag{2}$$

$$= 1 - P_{H_0}(Z \leq c - 1) \tag{3}$$

$$= 1 - \sum_{z=0}^{c-1} \binom{n}{z} p_0^z (1-p_0)^{n-z} \tag{4}$$

Die Nullhypothese wird abgelehnt, wenn $p(c)$ eine vorher festgelegtes Niveau α unterschreitet oder trifft, also wenn $p(c) \leq \alpha$.

3.2.3 Exakter Test nach Fisher

Der exakte Test nach Fisher (vgl. **Hartung**) für einseitige Hypothesen dient der Überprüfung, ob eine negative stochastische Abhängigkeit zwischen zwei dichotomen Merkmale A und B vorliegt. Bei einer anderen Skalierung kann eine Dichotomisierung - eine Einteilung in zwei disjunkte Klassen - vorgenommen werden, sodass der Test anwendbar ist. Es ist vorausgesetzt, dass beide Merkmale jeweils unabhängig identisch verteilt sind. Es liegen die Bezeichnungen der theoretischen Wahrscheinlichkeiten aus der Vierfeldertafel in Tabelle 3 zugrunde. Die zu überprüfende Hypothese lautet:

$$H_0 : p_{11} \leq p_{1\cdot} \cdot p_{\cdot 1} \quad \text{vs.} \quad H_1 : p_{11} > p_{1\cdot} \cdot p_{\cdot 1} \tag{5}$$

Unter der Nullhypothese sind A und B unabhängig verteilt und damit ist die Teststatistik Z hypergeometrisch Verteilt. Die Parameter der Verteilung ergeben sich aus den Randhäufigkeiten der Vierfeldertafel (s. Tab. 2). Dem Urnenmodell folgend, werden $n_{1\cdot}$ Kugeln aus einer Urne mit n_1 . weißen und n_2 . schwarzen Urnen gezogen. Der Wert der Teststatistik entspricht der Anzahl an gezogenen weißen Kugeln n_{11} . Der p -Wert berechnet sich demnach wie folgt:

$$P_{H_0}(Z = z) = \frac{\binom{n_{1\cdot}}{z} \binom{n_{2\cdot}}{n_{1\cdot}-z}}{\binom{N}{n_{1\cdot}}} \quad (6)$$

$$p(n_{11}) = P_{H_0}(Z \geq n_{11}) \quad (7)$$

$$= 1 - P_{H_0}(Z \leq n_{11} - 1) \quad (8)$$

$$= \sum_{k=0}^{n_{11}-1} P_{H_0}(Z = k) \quad (9)$$

Unterschreitet oder trifft der p -Wert ein vorher festgelegtes Niveau α , so wird die Nullhypothese abgelehnt. In diesem Fall kann davon ausgegangen werden, dass zwischen A und B eine positive stochastische Abhängigkeit vorhanden ist.

Tabelle 3: Vierfeldertafel mit theoretischen Wahrscheinlichkeiten p_{ij}

		A		
		1	0	Σ
B	1	p_{11}	p_{12}	$p_{1\cdot}$
	0	p_{21}	p_{22}	$p_{2\cdot}$
Σ		$p_{\cdot 1}$	$p_{\cdot 2}$	1

4 Durchführung

Das zu untersuchende Gebiet erstreckt sich in Nord-Süd Richtung von $51,56^\circ$ N bis $51,47^\circ$ N und in West-Ost-Richtung von $0,145^\circ$ W bis $0,005^\circ$ W. Dies entspricht etwa 10km Ausdehnung in beide Richtungen. Der Bereich wurde exemplarisch gewählt um den Stadtkern sowie Teile der Boroughs Tower Hamlets, Southwark, City of Westminster, Hackney und Islington abzudecken.

Das Einlesen der Daten in R erfolgt wie in Abschnitt 3.1.1 auf Seite 12 beschrieben und wird mit den dort angegebenen Programmen und dazugehörigen Kommandozeilenparametern (siehe 1 bis 3) ausgeführt.

Als Parameter für den Algorithmus und die Auswertung wird eine Laufweite von 1000 Metern gewählt und eine durchschnittliche Straßensegmentlänge von 50 Metern.

Die Verarbeitung der Daten und die Durchführung des Algorithmus wird wie in den Abschnitten 3.1.3 bis 3.1.4 durchgeführt.

Nun liegen für jedes Straßensegment Unfall-ID, sowie Entfernung in Laufweite und kumulierte Distanz für jeden Unfall innerhalb der Laufweite vor. Mit diesen Daten kann die statistische Auswertung für jedes Straßensegment einzeln vorgenommen werden.

4.1 Abweichungen in der Anzahl der in Unfällen verletzten Passanten

Die Unfälle in der inneren Umgebung sind jeweils solche, für die die kumulierte Distanz zwischen 0 und 500 Metern liegt. Unfälle in der äußeren Umgebung weisen kumulierte Distanzen zwischen 500 und 1000 Metern auf. Es wird für jede Umgebung die Anzahl der in Unfällen verwickelten Fußgänger gezählt. Dabei bezeichnet n_i die Anzahl der Unfälle in der inneren Umgebung und n_a die Anzahl in der äußeren Umgebung. Darauf aufbauend, wird der relative Anteil der Unfallopfer in der inneren Umgebung gegenüber der Unfallopfer in beiden Umgebungen $r = \frac{n^i}{n^i + n^a}$ berechnet. Der Wert r dient dann als Kennzahl zur Färbung des Straßensegmentes, sodass ein hoher ($r > 0.5$) relativer Anteil an Unfallopfern in der inneren Umgebung zu einer roten Färbung führt. Umgekehrt würde ein niedriger Anteil ($r < 0.5$) an Unfallopfern in der inneren Umgebung zu einer grünen Färbung führen. Da jedoch lediglich nur Unfallschwerpunkte von Interesse sind werden solche Fälle nicht eingezeichnet, wie im weiteren Verlauf ersichtlich wird.

Die zu überprüfende Annahme, dass die innere Umgebung keinen Unfallschwerpunkt darstellt, lässt sich gemäß der Vorüberlegung (vgl. Kap. ??) in eine quantitative Hypothese umformulieren. Diese besagt, dass in der inneren Umgebung nicht mehr Unfälle liegen als in der äußeren. Da beide Umgebungen die selbe Straßenstrecke überdecken ist diese Annahme gerechtfertigt. Demnach ist die Wahrscheinlichkeit p , dass ein Unfall in der inneren Umgebung liegt kleiner gleich der Wahrscheinlichkeit, dass ein Unfall in der äußeren Umgebung liegt. Somit lässt sich die zu überprüfende Hypothese wie folgt formulieren:

$$H_0 : p \leq 0.5 \quad \text{vs.} \quad H_1 : p > 0.5$$

Damit ist unter H_0 die Anzahl der $n^a + n^i$ Unfälle in den beiden Umgebungen Binomialverteilt mit dem Erwartungswert $\frac{n^a + n^i}{2}$. Zur Beurteilung, ob die Anzahl der Unfälle in der inneren Umgebung signifikant höher ist, kann folglich der einseitige

Binomialtest, wie in Kapitel 3.2.2 beschrieben, herangezogen werden. Der p -Wert eines Straßenabschnitts berechnet sich somit wie folgt:

$$p(n^i) = 2 \cdot 0.5^{n^i+n^a} \sum_{z=0}^{n^i} \binom{n^i + n^a}{z} \quad (10)$$

Der Binomialtest geht davon aus, dass die Ereignisse unabhängig voneinander sind. Gerade bei Verkehrsunfällen kann davon in der Realität nicht immer ausgegangen werden. Es kann jedoch angenommen werden, dass die Abhängigkeiten nur für Unfälle in einem sehr kurzen Zeitraum gegenüber den vorliegenden sechs Jahren besteht und damit zu vernachlässigen ist.

Dieses Testergebnis wird hier genutzt um zu entscheiden, ob die relative Abweichung r aus einer signifikanten Ungleichverteilung der Unfälle auf die äußere und innere Umgebung herrührt. Das hier gewählte Niveau α beträgt 0.1. Ein Straßensegment wird demnach erst in die Grafik eingezeichnet, wenn für die Auswertung einen p -Wert kleiner als 0.1 ermittelt wird. Ebenfalls sollen eindeutigere Testentscheidungen jedoch auch hervorgehoben werden. Dazu wird die Transparenz eines Straßensegments linear verringert, je kleiner der p -Wert ist. Ein p -Wert von 0 bedeutet keine Transparenz des Straßensegments. Das Ergebnis ist in Abbildung 7 zu sehen.

4.2 Unterschiede der Unfallzahlen in Abhängigkeit der Altersgruppe

Die Wahl der Umgebungen erfolgt über die Entfernung in Laufweite. Unfälle welche innerhalb von 500 Metern Laufweite liegen fallen in die innere Umgebung. In der äußeren Umgebung liegen Unfälle mit einer Laufweite zwischen 500 Metern und einem Kilometer. Für jeden, in einem Verkehrsunfall verletzten Fußgänger, wird seine Altersgruppe erfasst und die Beobachtung in die Klasse der 6 bis 15 jährigen oder der anderen Alters eingeteilt. Die Ergebnisse lassen sich für jedes Straßensegment in eine Vierfeldertafel (vgl. Tab. 4) eintragen.

Tabelle 4: Vierfeldertafel

		Alter		
		6 – 15 J.	andere	Σ
Umgebung	innere	n_{11}	n_{12}	$n_{1\cdot}$
	äußere	n_{21}	n_{22}	$n_{2\cdot}$
	Σ	$n_{\cdot 1}$	$n_{\cdot 2}$	N

Für die Färbung des Straßensegments wird die Risikodifferenz (s. Kap. 3.2.1) genutzt. Sind in der inneren Umgebung anteilig mehr 6 bis 15-jährige Kinder verletzt als in der

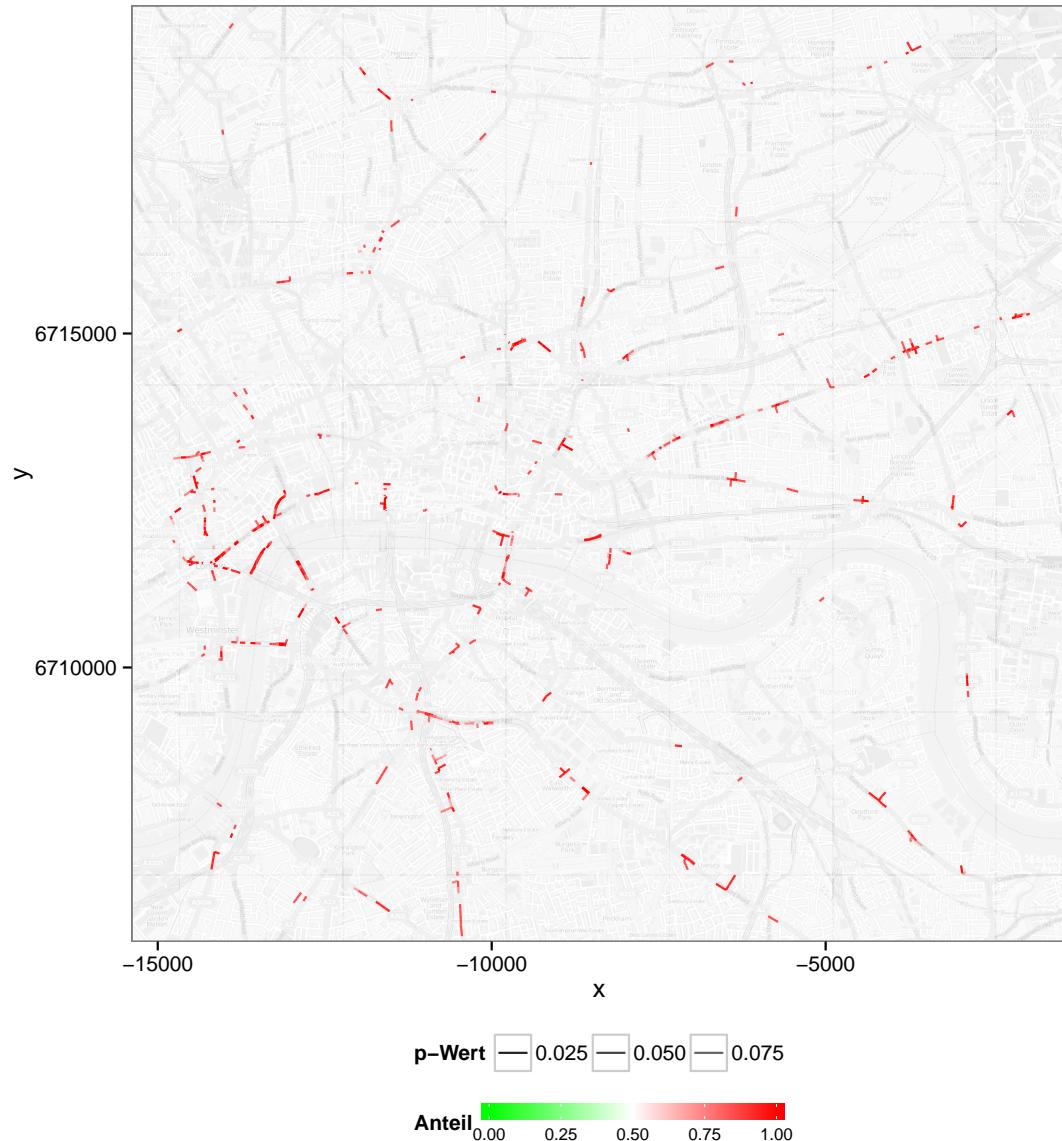


Abbildung 7: Diese Karte zeigt die, durch das in Kapitel 4.1 durchgeführte Verfahren, ermittelten Unfallschwerpunkte für Fußgänger im Straßenverkehr.

äußerer Umgebung, so nimmt die Kennzahl einen positiven Wert an. Die maximale Risikodifferenz kann höchstens den Wert 1 annehmen. Je näher der Wert an der 1 ist, desto intensiver die Rotfärbung des Straßensegments. Ein Wert nahe der 0 führt zu einer weißen Färbung. Wäre die Risikodifferenz negativ, so würde sie analog grün gefärbt. Da jedoch nur Unfallschwerpunkte von Interesse sind, werden diese Fälle nicht eingezeichnet. Dies wird im folgenden erläutert.

Auch hier wird die Annahme, dass für die Gruppe der 6 bis 15-jährigen Passanten in der inneren Umgebung ein Unfallschwerpunkt vorliegt der Vorüberlegung (vgl. Kap. ??) folgend in eine empirische Annahme umgeformt.

Es gilt zu überprüfen, ob für das jeweilige Straßensegment das Alter eines Unfallopfers in der Kategorie 6 bis 15-Jahren unabhängig ist von der Umgebung, in welcher der Unfall liegt. Diese Hypothese wird mit dem einseitigen Exakten Test nach Fisher (s. Kap. 3.2.3) überprüft. Der p -Wert berechnet sich dann wie folgt:

$$p(n_{11}) = \sum_{k=0}^{n_{1.}} P_{H_0}(Z = k) \cdot \mathbb{I}_{[0, P_{H_0}(Z=n_{11})]}(P_{H_0}(Z = k))$$

$$\text{mit } P_{H_0}(Z = z) = \frac{\binom{n_{1.}}{z} \binom{n_{2.}}{n_{1.}-z}}{\binom{N}{n_{1.}}}$$

Ähnlich wie im vorhergehenden Abschnitt wird auch hier das Testergebnis genutzt, um zu beurteilen, ob die Risikodifferenz auch mit einer signifikanten Abhängigkeit der Altersgruppe und der Umgebung einhergeht. Das Niveau α beträgt 0.1. Ist der errechnete p -Wert für ein Straßensegment kleiner α so wird dieses Straßensegment in die Karte eingezeichnet. Je eindeutiger das Testergebnis im Sinne des p -Wertes ist, desto intensiver (Verringerung der Transparenz) die Zeichnung des Straßensegments. Die so erzeugte Grafik ist in Abbildung 8 zu sehen.

5 Auswertung und Problematiken

Der verwendete Datensatz ist als Fallbeispiel für die vorgestellten Verfahren zu betrachten, darum wurde auf eine Auswertung diesbezüglich spezifischen Ergebnisse verzichtet. Diese sei Lesern mit geeignetem Hintergrundwissen überlassen.

Es ist beachtenswert, dass selbst der großzügig gewählte Zeitraum von sechs Jahren in dem dicht bewohnten Gebiet Londons auf vielen Straßen keine signifikanten Ergebnisse liefert. Dies liegt zum einen an relativ niedrigen Zahlen von Verkehrstöpfen in bestimmten Gebieten und den relativ klein gewählten Umgebungen. Die gewählten 500 Meter Straßüberdeckung je Umgebung führen im Schnitt zu in Laufweite relativ nah gelegenen Unfällen (vgl. Abb. 9). Der Einfluss der aufgeführten Störvariablen wie Verkehrsaufkommen und Fußgängerfrequentierung sind in diesem kleinen Umkreis als konstant anzusehen. Sie verfälschen die Resultate demnach nicht.

Eine kurze Rechnung zeigt, dass für das gewählte Niveau von $\alpha = 0.1$ mindestens 5 und für das Niveau $\alpha = 0.05$ mindestens 6 Unfälle in beiden Umgebungen zusammen liegen müssen um ein signifikantes Testergebnis zu erhalten. Abbildung 10 zeigt, dass in den meisten Umgebungen zu wenig Unfälle liegen, um diese Schwelle zu überschreiten. Gleichzeitig ist auch ersichtlich, dass die Fallzahlen für die Straßensegmente in einer Größenordnung sind, welche dem Exakten Test nach Fisher den Vorzug gegenüber dem Chi-Quadrat-Test geben.

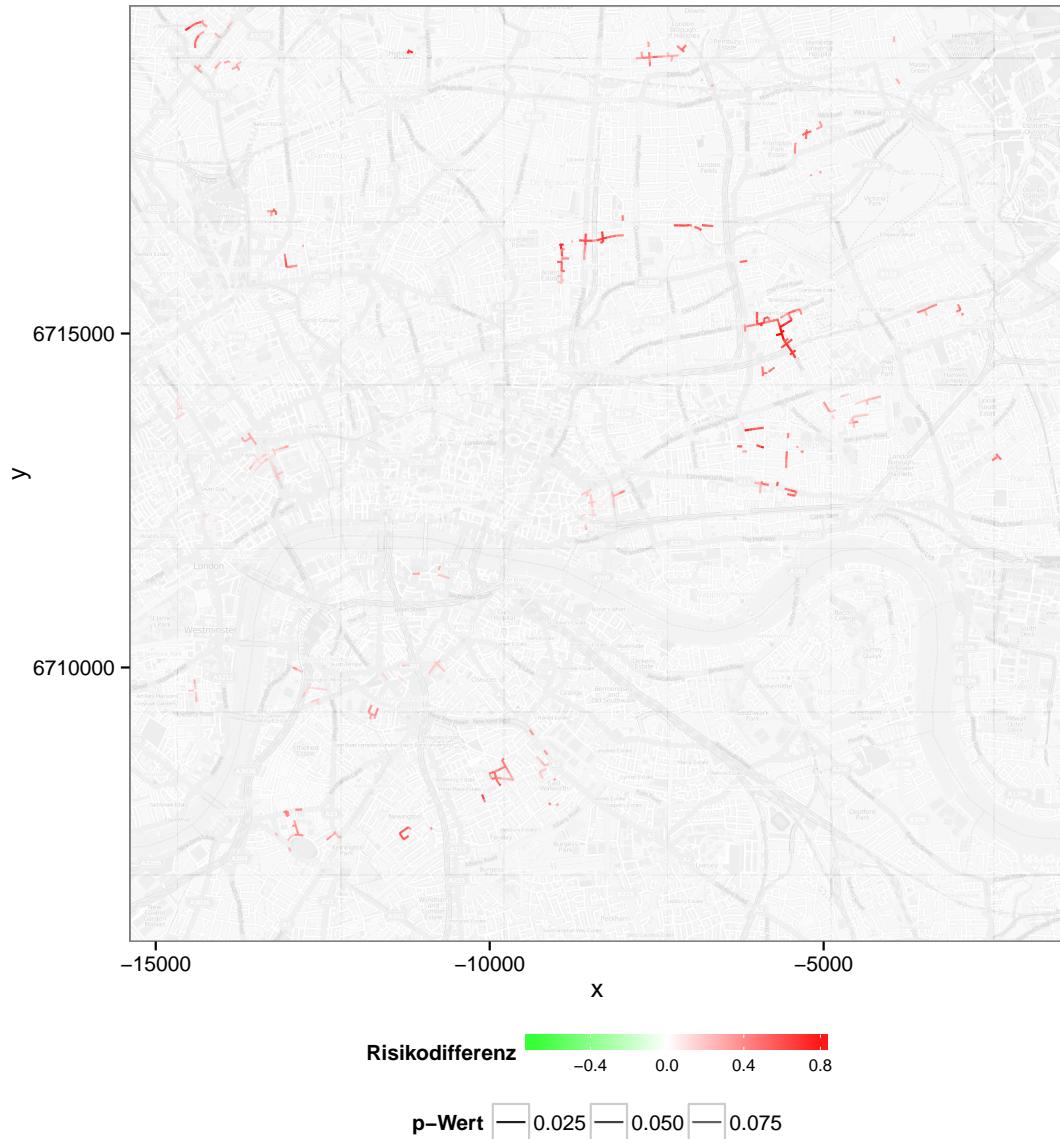


Abbildung 8: Diese Karte zeigt die, durch das in Kapitel 4.2 durchgeführte Verfahren, ermittelten Unfallschwerpunkte für Kinder im Alter von 6 bis 15 Jahren im Straßenverkehr.

Alternativ zu dem hier gewählten einseitigem Testproblem, kann auch ein zweiseitiges Testproblem gewählt werden. So würden auch Straßenabschnitte gezeigt werden können, welche vergleichsweise wenige Unfälle aufweisen.

Die Problematik des multiplen Testens wird vernachlässigt. Zwar sind die p -Werte von nebeneinanderliegenden Straßensegmenten hochkorreliert. Es kann argumentiert werden, dass durch die vielen Tests die auf immer nur leicht unterschiedlichen Umgebungen durch zufall eine Umgebung genau so gelegt ist, dass der p -Wert das Niveau unterschreitet. Dem wird jedoch entgegengewirkt, da die p -Werte für die benachbarten Straßensegmente ebenfalls ersichtlich sind, sodass der geschilderte Fall ersichtlich und eine Fehlinterpretation vermeidbar wird. Dem Betrachter sollte dies jedoch bewusst

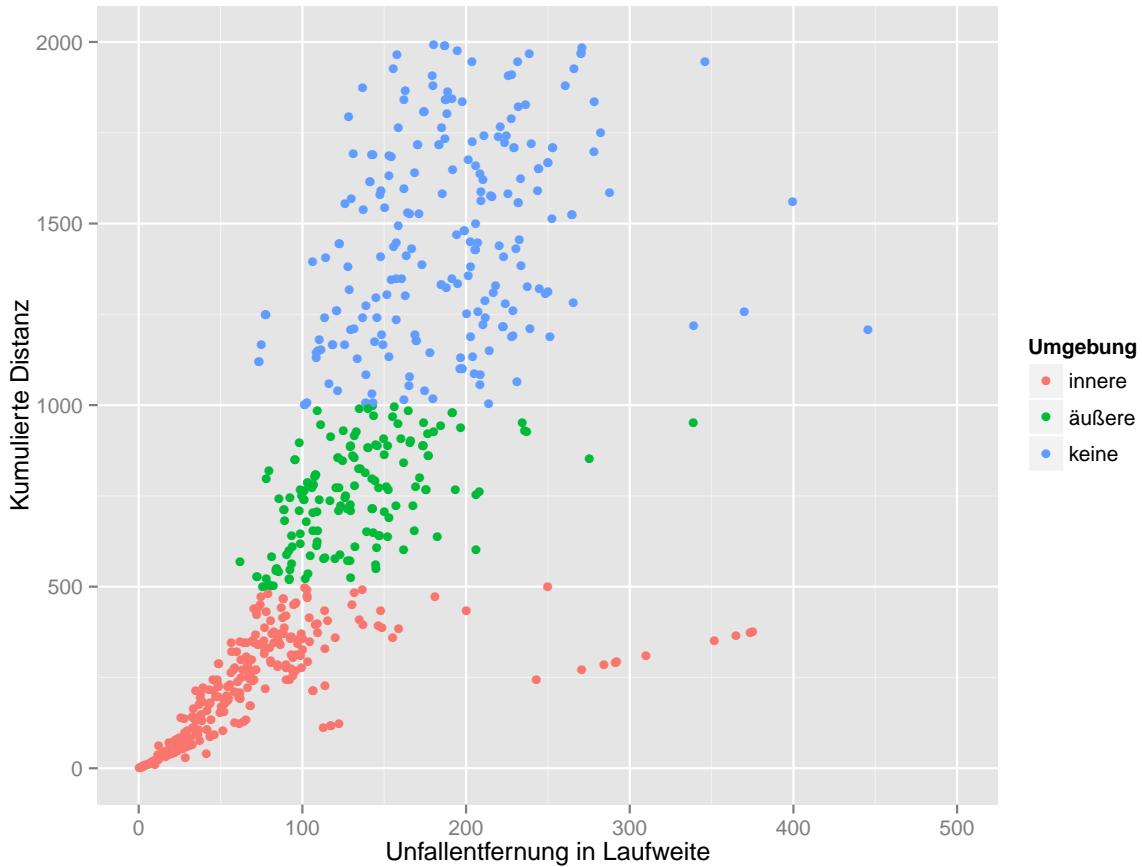


Abbildung 9: Ausschnitt aus den Ergebnissen der Distanzberechnung der Unfälle für eine Stichprobe von 500 Straßensegmenten. Es fehlen die Einträge mit größeren Werten auf den Achsen.

sein.

Die Karte in Abbildung 8, in welcher die Unfallzahlen in den Umgebungen der 6 bis 15 Jährigen mit denen anderen Alters verglichen werden, darf nicht zu der Fehlinterpretation führen, dass grün gefärbte Straßensegmente per se für Kinder sicherer sind als an anderen Orten weiß bis rot gefärbte Segmente. Es handelt sich lediglich um Aussagen die relativ zu den Gesamtunfallzahlen auf Basis der inneren und äußeren Umgebung getroffen werden. Ein Straßenabschnitt, in dem absolut mehr Kinder in Verkehrsunfälle verwickelt wurden, muss hier nicht als Unfallschwerpunkt dargestellt werden, als ein Straßenabschnitt mit absolut weniger in Verkehrsunfällen verwickelten Kindern. Gegebenenfalls kann die Auswertung für Untergruppen der Personen auch ohne Berücksichtigung der restlichen Personen geschehen, was zu einer Auswertung wie in Kapitel 4.1 führt.

Dem schließt sich ein Kritikpunkt an, dass die Wahl der Umgebung in der Darstellung nicht ersichtlich ist. Die Variante aus Kapitel 4.1 wendet die Laufweite zur Wahl der Umgebung und somit ist abschätzbar, welche Straßen in die Auswertung mit

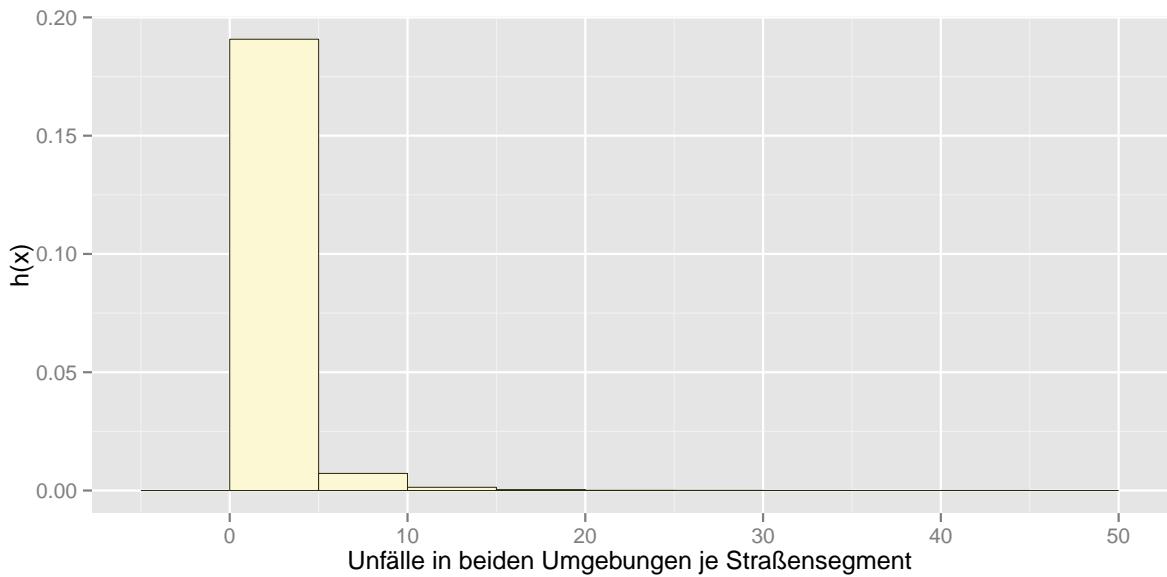


Abbildung 10: Ausschnitt aus den Ergebnissen der Distanzberechnung der Unfälle für eine Stichprobe von 500 Straßensegmenten. Es fehlen die Einträge mit größeren Werten auf den Achsen.

einbezogen werden. Hingegen bleibt dies bei der Auswertung in Kapitel 4.2 unersichtlich und im Nachhinein schwer einzuschätzen. Im Zweifel sollte ein Straßensegment und die Umgebung einzeln untersucht werden. Es sei angemerkt, dass die gewählte Auswertung Darstellung der Daten eher als Übersicht und Hilfsmittel angedacht ist.

Das Verfahren aus Kapitel 4.2 nutzt eine relativ große Umgebung. Diese wurde so groß gewählt, da die es vergleichsweise wenige Verkehrsunfälle mit Fußgängern in der Altersgruppe der 6 bis 15-jährigen gibt. Es kann geschehen, dass für ein zu untersuchendes Segment die Unfälle in der inneren Umgebung in entgegengesetzten Richtungen am Rand der Umgebung liegen. Die Straßensegmente auf denen die Unfälle jedoch tatsächlich liegen kein signifikantes Testergebnis liefern. Dies ist tatsächlich der Fall, wie Abbildung 11 zeigt.

In den vorliegenden Daten sind auch Informationen zur Schwere der Unfälle vorhanden. Diese wurden aus zwei Gründen hier nicht verwendet. Einerseits gibt keine genauen Angaben wie diese Klassierung vorgenommen wurde. Andererseits wird davon ausgegangen, dass jeder *leichte* Unfall durch andere Umstände, auf die durch straßenbauliche Maßnahmen kein Einfluss zu nehmen ist, auch zu einem schwererem Unfall hätte werden können. Gleichermaßen gilt umgekehrt für schwere Unfälle. Eine Ausnahme bildet jedoch die Geschwindigkeitsbegrenzung, welche unmittelbaren Einfluss auf die Schwere der Verletzung der Fußgänger hat¹

¹Impact Speed and a Pedestrian's Risk of Severe Injury or Death - AAA Foundation for Traffic Safety
<http://www.aaafoundation.org/pdf/2011PedestrianRiskVsSpeed.pdf>

Für die Wahl der Umgebungen könnte eine Berücksichtigung des Straßentyps und der Geschwindigkeitsbegrenzung nützlich sein. Diese Informationen liegen weitestgehend in den Daten der Open Street Map vor. So könnte gesichert werden, dass nur Straßen gleichen Typs miteinander verglichen werden. Dies würde dazu führen, dass Faktoren wie Verkehrsaufkommen und Geschwindigkeitsbegrenzung einen vermindernten Einfluss auf die Ergebnisse haben.

Die hier vorgestellte Analyse behandelt keine Unfälle welche innerhalb einer Entfernung von 20 Metern zu einer Kreuzung liegen, da Unfälle an Kreuzungen häufig andere Ursachen haben, als solche die auf gerader Strecke geschehen. Es ist denkbar diese ähnlich der vorgestellten Verfahren auszuwerten und die Kreuzungen als unterschiedlich große und gefärbte Punkte zusätzlich einzuteichnen. Dies würde jedoch nur für Kartenausschnitte im größeren Maßstab die Übersichtlichkeit nicht beeinträchtigen.

6 Zusammenfassung

Die Digitalisierung der Gesellschaft bringt eine Datenfülle mit sich, welche stets nach neuen und verbesserten Methoden verlangt diese auszuwerten. Eine besondere Klasse unter diesen Daten sind Geodaten. Beobachtungen mit Breiten und Längengrad verknüpft sind Grundlage der *Geospatial Analysis*. Im Zuge der *Open Data*-Bewegung sind immer mehr Daten öffentlich zugänglich. Diese Trends motivieren diese Arbeit. Die britische Regierung veröffentlicht auf data.gov.uk unter anderem die Daten aller von der Polizei erfassten Unfälle in den Jahren 2005 bis 2011. Jede Beobachtung in diesem Datensatz enthält neben weiteren Merkmalen Informationen zu dem Ort des Geschehens und Alter der Unfallopfer. Von Interesse sind jene Unfälle, welche nicht in unmittelbarer Nähe einer Kreuzung liegen und in welchen Fußgänger verletzt wurden. Die Angaben über Breiten und Längengrad werden verwendet um jedem Unfall eine Position auf einer Straße zuzuweisen. Die Straßeninformationen werden der *Open Street Map* entnommen und mit dem R-Paket *osmar* eingelesen. Die Straßen werden in kurze Straßensegmente (ca. 50m) unterteilt. Der Mittelpunkt jedes Straßensegments ist Grundlage für die weitere Untersuchung. Für jeden Mittelpunkt werden die Unfälle, welche in einer Laufweite von bis zu 1000 Metern liegen, mitsamt der entsprechenden Entfernung ermittelt. Ebenfalls wird ein Distanzmaß eingeführt, welches für jede Laufdistanz vom Mittelpunkt zu einem Unfall die Strecke angibt, welche bis zu dem Unfall zurückgelegt wird, würde auf allen Abzweigungen parallel gegangen werden. Dies gibt die Möglichkeit die Unfälle auf eine innere und äußere Umgebung einzuteilen, welche jeweils die gleiche Straßenstrecke überdecken. Mit diesen gleich großen Umgebungen von jeweils 500 Metern überdeckter

Straßenstrecke kann davon ausgegangen werden, dass ebenfalls die Anzahl der Unfälle in beiden Umgebungen gleich groß ist. Durch Anwendung des Binomialtests kann festgestellt werden, ob die Anzahl in der inneren Umgebung signifikant höher ist. Der relative Anteil der Unfälle in der inneren Umgebung von den Unfällen in beiden Umgebungen dient dabei als Kennzahl. Die Kennzahl zusammen mit dem Testergebnis geben Indizien dafür, ob die innere Umgebung Unfallschwerpunkt ist.

Daneben wird in Kapitel 4.2 eine Möglichkeit vorgestellt die Beurteilung für eine bestimmte Personengruppe zu treffen. In dieser Arbeit wird beispielhaft die Gruppe der 6 bis 15 jährigen Kinder von Andersaltrigen unterschieden. Die Unfälle werden anhand ihrer Entfernung in Laufweite von Mittelpunkt des Straßensegments in die äußere und innere Umgebung eingeteilt. In die innere Umgebung fallen Unfälle mit einer Entfernung bis zu 500 Metern. Weiter, bis zu einem Kilometer, entfernte Unfälle werden in die äußere Umgebung eingeteilt. In der Annahme, dass es keinen Unfallschwerpunkt für die spezifische Personengruppe gibt, ist die Anzahl der Unfälle unabhängig von der Umgebung und der Personengruppe. Dem zu folge kann der *Exakte Test nach Fisher* verwendet werden um zu entscheiden, ob eine signifikant positive Abhängigkeit zwischen dem Merkmal „innere Umgebung“ und „Alter: 6 bis 15 Jahre“ vorliegt. Als Kennzahl für den Unterschied der Unfallanteile in den Umgebungen für die Personengruppe wird die Risikodifferenz verwendet. Das Testergebnis und diese Kennzahl indizieren mögliche Unfallschwerpunkte für die bestimmte Personengruppe.

Die Ergebnisse aus beiden Verfahren werden in einer Karte dargestellt, in welcher die Straßensegmente anhand der ermittelten Kennzahl gefärbt werden. Die Straßensegmente werden eingezeichnet, wenn der p -Wert des Tests kleiner gleich dem gewählten Niveau $\alpha = 0.1$ ist.

Sich durch diese Verfahren ergebenden Problematiken werden in Kapitel ?? besprochen. Unter anderem sind dies die korrelierten Testergebnisse nebeneinanderliegender Straßensegmente, mögliche Fehlinterpretationen der erstellten Karten und Unzulänglichkeiten dieser Darstellungsvariante.

In weiterer Entwicklung könnte die Unfallschwere Einfluss mit in die Analyse finden. Unterschiedlich klassifizierte Straßen (Hauptstraßen, Straßen in Wohngebieten) sollten gegebenenfalls aufgrund von unterschiedlichem Verkehrsaufkommen nicht in den Umgebungen gemischt werden. Eine weitere Analyse der Unfallzahlen an den Kreuzungen und ein Vergleich mit benachbarten Kreuzungen könnte ähnlich dem vorgestellten Verfahren durchgeführt werden und Einzug in die Grafik erhalten.

Ebenfalls anzumerken ist, dass das vorgestellte Verfahren sich für viele weitere ortsbezogene Informationen verwenden lässt. So ist es denkbar Informationen zu Straftaten auf

eine ähnliche Weise auszuwerten. Es ließen sich Straßenzüge finden, welche eine besonders hohe Einbruchquote aufweisen. Auch für geschäftliche Interessen wäre es etwa für eine Neueröffnung einer Verkaufsstelle von Interesse wo in Laufweite wenige Geschäfte der ähnlichen Kategorie liegen, gleichzeitig jedoch ein Indikator für das Aufkommen an potenziellen Kunden auf einem guten Niveau liegt.

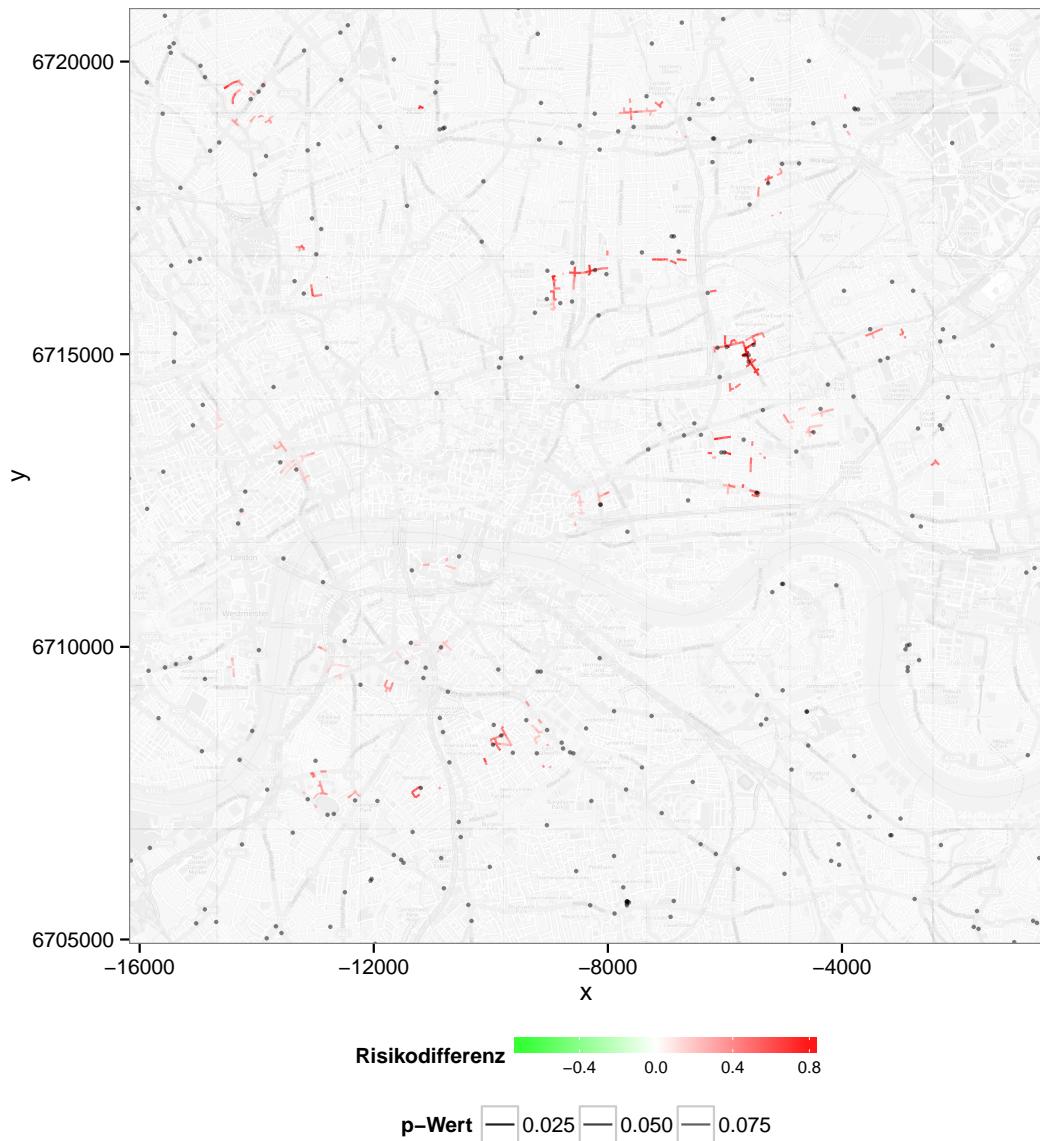


Abbildung 11: Diese Karte zeigt die Abbildung 8. Zusätzlich sind die Unfallpunkte dieser Altersgruppe eingezeichnet. Es ist zu sehen, dass Unfallschwerpunkte an Stellen markiert wurden, welche tatsächlich jedoch keine Unfälle aufweisen.

Literatur

Downloads von OSM-Dateien. Runtergeladen am 18.09.2012. Cloudmade. URL: http://downloads.cloudmade.com/europe/northern_europe/united_kingdom/united_kingdom.highway.osm.bz2.

Eugster, Manuel J. A. und Thomas Schlesinger (2010). „osmar: OpenStreetMap and R“. In: *R Journal*. Accepted for publication on 2012-08-14. URL: <http://osmar.r-forge.r-project.org/RJpreprint.pdf>.

Fellows, Ian und using the JMapView library by Jan Peter Stotz (2012). *OpenStreetMap: Access to open street map raster images*. R package version 0.2. URL: <http://CRAN.R-project.org/package=OpenStreetMap>.

Keitt, Timothy H., Roger Bivand, Edzer Pebesma und Barry Rowlingson (2012). *rgdal: Bindings for the Geospatial Data Abstraction Library*. R package version 0.7-22. URL: <http://CRAN.R-project.org/package=rgdal>.

Osmfilter - OpenStreetMap Wiki. Heruntergeladen am 1.7.2012. URL: <http://wiki.openstreetmap.org/wiki/Osmfilter>.

Osmosis - OpenStreetMap Wiki. Heruntergeladen am 1.7.2012. URL: <http://wiki.openstreetmap.org/wiki/Osmosis>.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria. URL: <http://www.R-project.org/>.

Rüger, Bernhard (2002). *Test- und Schätztheorie*. Bd. Band II: Statistische Tests. München: Oldenbourg. ISBN: 3-486-25130-9.

Wickham, Hadley (2009). *ggplot2: elegant graphics for data analysis*. Springer New York. ISBN: 978-0-387-98140-6. URL: <http://had.co.nz/ggplot2/book>.

Mit meiner Unterschrift versichere ich, dass ich diese Arbeit eigenständig verfasst, sowie alle Zitate korrekt gekennzeichnet habe.

Jakob Richter

A Anhang

A.1 R-Code