Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Hausarbeit

Eduard Weigandt

Visual Analytics

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Eduard Weigandt

Visual Analytics

Hausarbeit eingereicht im Rahmen der Vorlesung AW2

im Studiengang Master of Science Informatik
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuende Prüfer: Prof. Dr. Kai von Luck & Prof. Dr. Bettina Buth

Eingereicht am: 31. August 2014

Eduard Weigandt

Thema der Arbeit
Visual Analytics

Stichworte
Visualisierung, Visual Analytics, Konzepte, Zeitreihenanalyse, Box-Jenkins-Methode, Räumliche-Analyse, Spatio-temporal-Analyse

Kurzzusammenfassung
In der hier präsentierten Recherche werden drei ausgesuchte Konzepte aus dem Bereich Visual Analytics vorgestellt. Dafür wurden drei Umsetzungen dieser Konzepte angeschaut und kritisch untereinander bewertet, um die Anforderungen an das Gebiet besser zu verstehen und Mittel kennen zu lernen, die es einem ermöglichen eigene Projekte umzusetzen. Die erste Methode setzt auf zeitlichen Daten auf und verwendet dabei die Box-Jenkins-Methode zur Bestimmung eines oder mehrerer Modelle, die die Daten angemessen repräsentieren, um anhand dieser dann Vorhersagen zu treffen. Im Fokus steht dabei die Integration des Verfahrens mit interaktiven Visualisierungen. Die zweite Umsetzung befasst sich mit einer effektiven Visualisierung von zusammenhängenden räumlichen Daten durch das Verwenden der Clustering Methode der *α-shapes* in Verbindung mit einer Heuristik, die zudem zum Anzeigen von Hierarchieebenen genutzt wird. Zum Schluss wird die Kombination aus beiden Arten von Daten in Ihren Möglichkeiten durch Korrelationskoeffizient am Beispiel von Kriminaldaten untersucht.

Inhaltsverzeichnis

1 Einleitung

In der ersten Ausarbeitung zum Thema *Visual Analytics* wurde eine Übersicht über das Fachgebiet gegeben, sowie die Motivation dafür erläutert. Die nun folgende Ausarbeitung zielt auf einen tieferen Einblick in drei wichtige Konzepte aus dem Bereich *Visual Analytics*. Zu diesem Zweck wurden drei aktuelle Arbeiten ausgewählt, die sich gut untereinander beurteilen lassen und ergänzen.

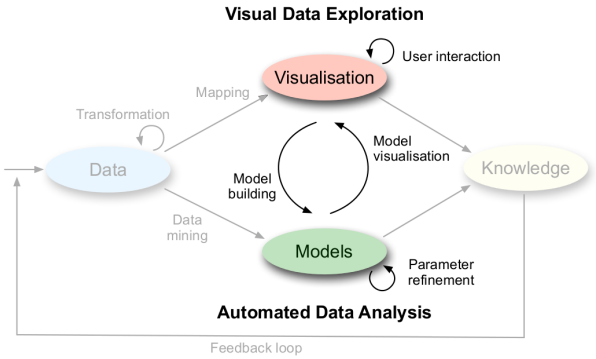


Abbildung 1.1: Visual Analytics Prozess und Komponenten

In Abbildung 1.1 sieht man zwei Komponenten, in denen die Konzepte Verwendung finden, hervorgehoben. Im Prozess der visuellen Analyse werden die unterschiedlichsten Arten von Daten durch diverse Methoden verarbeitet, um geeignete Modelle oder bessere Darstellungen von Daten zu erstellen. Aufbauend auf diesen Ergebnissen kann man dann mit Hilfe eines Experten eine Analyse zur Erfüllung einer Aufgabe oder das Entdecken von neuem Wissen ausführen. Für die Komponente die sich mit dem Modell der Daten auseinandersetzt wird in dieser Ausarbeitung der Aspekt der Zeit näher beleuchtet. Anschließend wird für die Visualisierung eine Methode vorgestellt, die zur Analyse von räumlichen Daten interessant ist und sich in das Umfeld des *Clustering* einordnet. Zum Schluss werden die neuen Möglichkeiten und Blickwinkel, die sich aus der Verbindung von zeitlichen und räumlichen Daten ergeben untersucht. Diese Beispiele zeigen hervorragend wie unterschiedliche Methoden bei *Visual Analytics* miteinander kombiniert werden können. Dabei wird auch gezeigt inwieweit diese Konzepte wichtig für die Umsetzung eigener Projekte zum Thema *Visual Analytics* sind und welche Anforderungen bei der Implementierung dieser anfallen.

2 Konzepte

2.1 Zeitreihenanalyse

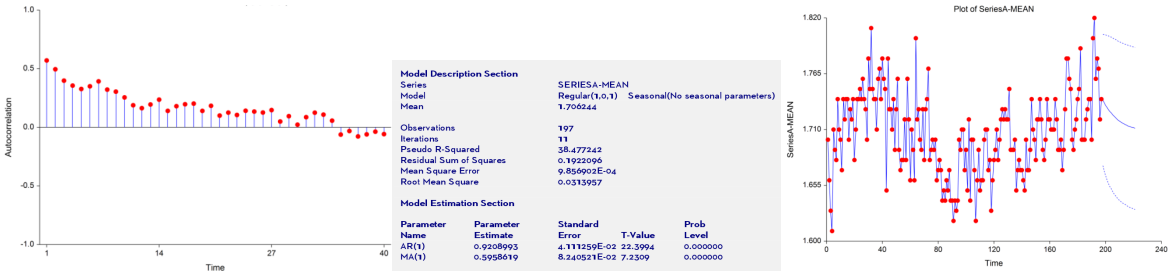
2.1.1 Aspekt der Zeit

In der Arbeit von Keim u. a. (2010), welche die Grundlagen für die folgenden Eigenschaften liefert, ist die Komplexität der Zeit gut erklärt. So z. B. besitzt die Zeit in Vergleich zu anderen Datenarten eine semantische Struktur und ist nicht einfach nur ”flach”. Unter einer semantischen Struktur versteht man z.B. die Hierarchie durch die einzelnen Untergliederungen in Sekunden, Minuten oder auch Wochen. Das sind nur ein paar der möglichen Arten die Zeit auszudrücken. Dazu kommt noch, dass diese Einheiten in unterschiedlichen Kalendersystemen vorkommen können. Daraus ergeben sich viele neue Wege wie man bestimmte Ereignisse im Auftreten untereinander untersuchen kann. Dabei ist interessant, dass man die Zeit in drei verschiedene Typen von Strukturen einteilen kann. Die erste ist die *ordered time* bei der Ereignisse linear nach einander ablaufen können oder sich zyklisch[1] wiederholen. Eine weitere Art ist die *branching time* unter welcher man die Beschreibung und den Vergleich zwischen alternativen denkbaren Szenarien versteht. Diese Art ist interessant für Felder, die sich mit der Planung oder Vorhersage beschäftigen. Die letzte Variante bildet die *multiple perspective time*, die den Blickwinkel auf eine Beobachtung im Mittelpunkt stellt. So können sich die Informationen[2] über ein und dasselbe Ereignis je nach Perspektive unterscheiden. Die letzteren beiden Typen von zeitlichen Strukturen werden relativ selten von *Visual Analytics* Verfahren oder Methoden aufgegriffen und stellen damit einen zukünftigen Ausblick zur weiteren Erforschung dar.

2.1.2 Box-Jenkins-Methode

Die erste Arbeit, die vorgestellt wird setzt sich mit der *ordered time* auseinander. Dabei werden für das Beispielszenario die Zeitreihen[3] von Patienten mit Herz- und Blutgefäßerkrankungen zur Vorhersage statistisch untersucht. Anhand der Ergebnisse aus diesem Beispielszenario könnte man dann ggf. vorbeugende Gegenmaßnahmen einleiten, wenn man einen zukünftigen Anstieg an Erkrankungen berechnet. Um eine gute Berechnung zu erstellen braucht man ein angemessenes Modell, welches die Eigenschaften der Daten widerspiegelt. Die Autoren Bogl u. a. (2013) haben zu diesem Zweck die Box-Jenkins-Methode ausgesucht.

[1] Frühling, Sommer, Herbst, ...
[2] z.B. Zeugenaussagen bei Verbrechen
[3] Ein Begriff aus der Statistik für eine zeitabhängige Folge von Datenpunkten. (Wikipedia)



(a) Autokorrelation (b) Parameter Schätzung (c) Daten und Vorhersage

Abbildung 2.1: Quelle: `http://www.ncss.com/` (27.08.2014)

Die Box-Jenkins-Methode wurde um 1970 herum von Box u. a. (2013) vorgestellt. Der Prozess in seiner Gesamtheit besteht aus drei iterativen Schritten, die man ausführen muss um ein angemessenes Modell in Form einer Polynominalfunktion für die Daten zu erhalten. Das Verfahren an sich ist sehr gut erforscht und stellt keine Neuerung auf dem Gebiet dar. Jedoch ist diese Implementierung die erste seiner Art, die die Techniken von *Visual Analytics* unter dem genannten Szenario nutzt. Bei dem erwähnten Modell handelt es sich in diesem Fall, um das ***Auto**regressive-Moving **A**verage* (ARIMA)[4]. Alleine gesehen ist diese Modellart nur für Daten ohne einen sichtbaren Trend einsetzbar. Dies wird Umgangen in dem man diesen im Bedarf raus rechnet.

Vorgehen. In Abbildung 2.1a sieht man den ersten Schritt. Das Ziel dieses Schrittes ist es anhand von Informationen über die Daten oder durch Funktionen zur Autokorrelation[5] eine Klasse von *ARIMA* Modellen auszuwählen, die man später an die Daten anpassen kann. Dabei versucht man das *Gesetz der Sparsamkeit*[6] einzuhalten. Durch diese erste Einschränkung kann man die Auswahl an Parametern und deren Werten für die endgültige Funktion verkleinern. Als zweiten Schritt wird eine Schätzung durchgeführt, bei der erste Werte für die Parameter gewählt werden (siehe Abbildung 2.1b). Diese Werte werden näherungsweise durch unterschiedliche Techniken berechnet, in diesem Fall bevorzugen die Autoren die *maximum likelihood-estimation* Methode. Der letzte Schritt beim Verfahren beinhaltet die Prüfung der restlichen Werte in der Zeitreihe, sowie das Finden von Fehlern. Für diese Untersuchung werden jeweils die Graphen in Abbildung 2.2 von **4a** bis **d** genutzt. Dabei dient jeder einzelne Graph, jeweils für einen anderen Aspekt der Zeitreihenanalyse. Das endgültige Ergebnis des ganzen Verfahrens sieht man in

[4] ARIMA bezeichnen lineare Modelle für stationäre, zeitdiskrete stochastische Prozesse. (Wikipedia)
[5] Die Wechselbeziehungen zwischen der verschobenen Variante eines zeitlichen Datenpunktes.
[6] Dieses besagt das man das einfachere Modell aus allen möglichen wählen sollte, da es nur ausreichend gute gibt, aber keine perfekt passenden. Je weniger Parameter das gewählte Modell benutzt, desto einfacher ist dieses.

Abbildung 2.1c, wo man die berechnete Vorhersage und den Trend zukünftiger Daten erkennen kann.

2.1.3 Bewertung

Der Schwerpunkt bei der Umsetzung war es einen besseren Arbeitsfluss für Analysten von Zeitreihenanalysen zu schaffen. Entscheidend dabei war die Kombination von automatischen Analysetechniken mit einer interaktiven Visualisierungen. Beim letzteren wird sichergestellt, dass der Analyst jeder Zeit in den Prozess eingreifen kann, um eigene Einstellungen vorzunehmen. Zu dem sollen die Visualisierungen der einzelnen Kriterien, die bei Abschnitt 2.1.2 genannt sind, den Experten im ganzen Prozess unterstützen. So kann man z.B. alle gewählten Modelle untereinander interaktiv Vergleichen, was hilfreich bei der Auswahl der Kriterien ist.

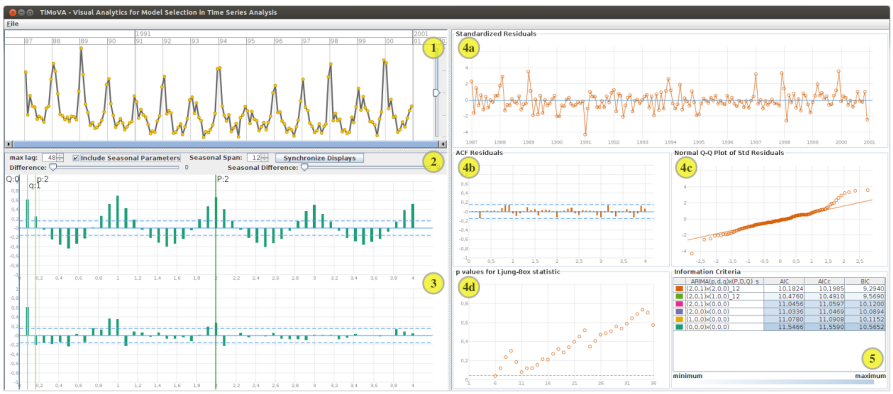


Abbildung 2.2: TiMoVa Applikation

Bei der Umsetzung von Bogl u. a. (2013) ist nicht die gewählte Methode, was besonders herausticht, sondern eher das Vorgehen bei der Erstellung und die detaillierte Dokumentation der wissenschaftlichen Arbeit. Bei den benutzten Daten gab es keinen Aufwand in der Transformation der Daten oder das Fehlen von Messungen, die die Auswertung erschwerten. Die Art der Daten war zudem eindimensional, also ohne weitere Variablen einbezogen. Bemerkenswert ist es jedoch, dass das Team mit Domänexperten[7] zusammengearbeitet hat, um eine direkte Rückmeldung zu erhalten. Somit wurde es sichergestellt, dass die Umsetzung auch richtig evaluiert werden konnte. Dies zeigt sich auch im Detail durch das genaue Aufzeigen des darunterliegenden *Visual Analytics* Prozesses. Es wurde zudem gezeigt, wie und wann die Applikation dem Benutzer beim Verfahren hilft. Das beschriebene Vorgehen und die Entscheidungen kann man für eigene Projekte einbeziehen. Die angestrebte Benutzerfreundlichkeit wurde aus meiner Sicht damit gut umgesetzt. Das was jedoch auffällt ist der Mangel an Experimenten in der Darstel-

[7] Visual Analytics und Statistik

lung, was jedoch nach den Autoren gewollt war, um einen Statistiker eine bekannte Umgebung zu präsentieren und den schnellen Start zu fördern.

2.2 Räumliche Analyse

2.2.1 Aspekt des Raums

Eine weitere Art von wichtigen Daten stellen die geographischen bzw. räumlichen dar. Nach Keim u. a. (2010) kann man z.B. mit geographischen Karten die Informationen aus der realen Welt aufzeigen, zusammenführen und erforschen, weil Karten eine abstrakte Sichtweise auf die komplexere Welt sind. Karten verdichten zudem essentielle geographische Attribute und Charakteristika. Jedoch muss vieles beachtet werden, um die richtigen Schlüsse zu ziehen. So z.B. stehen alle Dinge miteinander in Beziehung, aber die die näher bei einander sind um so mehr. Es gibt aber auch Ausnahmen, wenn z.B. zwei Dörfer von einem Gebirge getrennt sind, dann werden diese eher in keiner Beziehung zueinander stehen. Zu dem kann man die Entfernung zu zwei Punkten auf unterschiedliche Weise interpretieren. Die Distanz kann also trügerisch sein, je nachdem ob eine Straße oder die Luftlinie entscheidend ist. Erschwerend kommt hinzu, dass durch die Heterogenität in den dargestellten Informationsarten[8] man die passende Analysemethode wählen muss. Bei dem Szenario eines Chemieunfalls auf dem Land werden die Auswirkungen andere Ausmaße annehmen als auf einem Fluss. Räumliche Daten haben zu dem genauso wie zeitliche Informationen unterschiedliche Skalierungen, ob es die Verkehrsdichte oder die Richtung in die der Verkehr fließt ist, um ein Beispiel zu nennen. Dabei muss man sowohl den Maßstab als auch die Visualisierung beachten.

2.2.2 α-shapes / k-order

In der wissenschaftlichen Arbeit von Packer u. a. (2013) tauchen die oben genannten Aspekte und Fragestellungen zu räumlichen Daten ebenfalls auf. Das gewählte Realwelt-Szenario in dieser Ausarbeitung bestand aus der Untersuchung einer Stadt und deren öffentlichem Verkehrsnetz. Dabei wurde versucht Konzentrationspunkte (Autostopps) für die Auswertung besser sichtbar zu machen. Diese Ergebnisse sind für Stadtplaner von großem Wert, um den Verkehrsfluss von Autos durch das Verbessern von Verkehrsnetzen zu steuern.

α-shapes. Das endgültige Verfahren erlaubt es dem Benutzer die Parameter des α-shapes Algorithmus jederzeit anzupassen. Dabei wird der Analyst von der verwendeten Heuristik zu potentiell besseren Ergebnissen geleitet. Zu dem werden Hierarchieebenen durch das Darstellen von gröberen Regionen, die alle durch andere Farben markiert sind (siehe Abbildung 2.3), sichtbar

[8] Land, Höhenunterschiede, Wasser, etc.

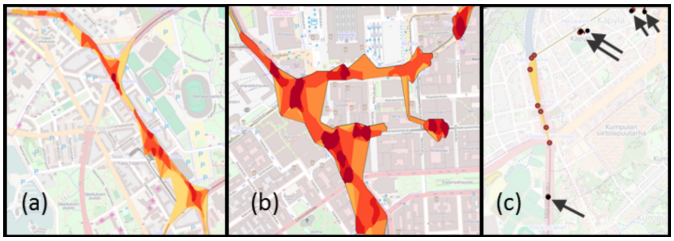
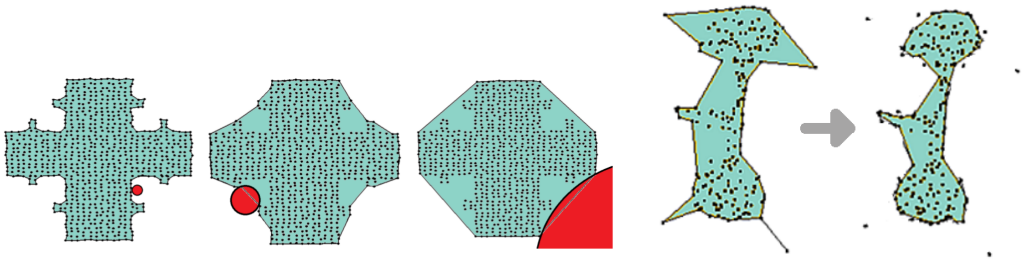


Abbildung 2.3: Ergebnis der Analyse von Busstopps in einer Stadt

gemacht. Die Funktionsweise vom α-shapes Algorithmus ist schnell erklärt. Es werden unterschiedliche α Werte gewählt, diese beschreiben jeweils den Radius eines Kreises. Der so erstellte Kreis mit dem Radius α wird an die Messpunkte angelegt, wobei die Punkte an der Kreisgrenze ein Polygon bilden. Der Kreis darf deswegen keinen Messpunkte überdecken (siehe Abbildung 2.4a) und alle so gefundenen Punkte fallen somit in ein und dieselbe Region. Am Anfang des Verfahrens gibt es eine leere Menge, die dann am Schluss alle Punkte in Form von Kanten zwischen zwei Punkten beinhaltet. Die einzelnen Schritte bis zur letzten Form sind mögliche Cluster. Ein großer Nachteil solcher Methoden besteht bei größeren Ausreißern in den Daten, da so unnötig große Cluster gebildet werden, die die Information verfälschen.



(a) $\alpha - shapes$ mit unterschiedlichem Parameterwert (b) Ausreißer Eliminierung

Abbildung 2.4

k-order α-shape. Um die Ausreißer aus den Cluster heraus zu rechnen, benutzen die Autoren den *k-order α-shape* Algorithmus von Krasnoshchekov u. Polishchuk (2008), welcher anhand eines selbstgewählten Schwellenwertes unterschiedliche Cluster vorschlägt. In Abbildung 2.4b sieht man rechts die bereinigte Variante des rechten Clusters als ein Beispiel.

Heuristik. Eine weitere wichtige Eigenschaft der Arbeit beinhaltet das semi-automatische Erstellung von unterschiedlich geformten Clustern, die zur hierarchischen Analyse genutzt werden können. Da bei dem Verfahren alle möglichen Formen berechnet werden, muss man sich nur die potentiell interessanten raus suchen. Gegebenenfalls kann man durch die Auswahl an Hierarchie-

ebenen neue Zusammenhänge unter den Daten entdecken. Diese Formen werden als *Breakpoints* bezeichnet und durch die folgende Heuristik beschrieben:

- Isolierte Cluster aus einem Punkt werden nicht betrachtet.
- Der Knotengrad darf nur 0 oder 2 sein $\rightarrow$ Löcher in Daten erkennen
- Jeder neu untersuchte Punkt muss in einem vorhanden Polygon enthalten sein
- Vorschläge durch Topologische Äquivalenzklassenbildung $\rightarrow$ straffe Formen ohne Löcher bevorzugen / so wenig Löcher wie möglich / Form mit max. α Wert

2.2.3 Bewertung

Die Autoren zeigen klar auf, wie sie die α-shapes einsetzen und anpassen, um eine gute Integration in den Analyseprozess zu erhalten. Dabei verwenden sie Erkenntnisse von anderen Autoren für die Eliminierung von Ausreißern. Eine vollständige Dokumentation[9] von der genauen Verwendung ist jedoch nicht mit angegeben.
Das Verfahren an sich hilft dem Analysten durch das Anzeigen von Hierarchien und den damit verbunden Parametervorschlägen bei der Erfüllung seiner Aufgaben. Die Parameterwerte kann der Benutzer zu jederzeit anpassen. Die Autoren haben leider keine Ergebnisse zu den Vergleichen mit anderen Heuristiken veröffentlicht, was für eine bessere Einsicht in das Thema hilfreich gewesen wäre. Im Gegensatz dazu wird eine gute kompakte Einführung in die verwandten Themengebiete, die sich mit den α-shapes beschäftigen, gegeben. Das verwendete Verfahren und die Berechnungen sind ebenso gut detailliert dokumentiert so dass man den Algorithmus für eigene Projekte einfach aufgreifen kann. Des Weiteren ist die Komplexität des Verfahrens zur α-shape-Berechnung linear und für alle anderen eingesetzter Methoden jeweils **O**(n log n). Es werden leider auch hier keine weiteren Vergleiche zu anderen Methoden gegeben, jedoch wird betont das der Beitrag in dieser Arbeit eine berechnungseffiziente Methode darstellt. Die Autoren sagen jedoch am Ende der Arbeit, dass das Verfahren folgenden schwächen aufweist:

- Es wurde kein Vergleich zu andern Clustering Methoden gemacht
- Es gibt effizientere Algorithmen um Ausreißer zu entfernen.
- Usability kann man noch weiter ausbauen, um Heuristiken einfach auszutauschen.
- Problemstellung wurde zu eng gefasst.

Die oben aufgezählten Schwächen bilden gute Ansatzpunkte für eigene Untersuchungen in diesem Gebiet und sollten beachtet werden.

[9] Es wird darauf hingewiesen, dass diese in einer anderen Ausarbeitung nachgeliefert wird.

2.3 Spatio-temporal Analyse

2.3.1 Verbindung von Raum und Zeit

In diesem letzten Abschnitt 2.3 werden die noch ausstehenden Anforderungen aus den ersten beiden Abschnitten behandelt, sowie der Vorteil in der Verbindung von Zeit und Raum in einer gemeinsamen Analyse. Eines der ersten Dinge die in Keim u. a. (2010) genannt werden ist der Nutzen in der Suche nach der Kausalität von Ereignissen. Dabei wird das Beispielszenario eines Risikoanalysten entworfen, der sich zukünftige Wetterdaten anschaut um daraus Schlüsse zu ziehen. Wenn man dadurch an Kenntnisse von einem Unwetter im Bereich einer Autobahn kommt, kann man als Versicherungsfirma dem entsprechend reagieren um Kosten im Automobilbereich einzusparen. Dabei ist es wichtig zu sagen, dass jede Beobachtung stark von den Daten abhängig ist und vielleicht nicht übertragbar ist. Nach demselben Prinzip wie: ”Wenn es regnet, dann ist die Straße nass.” Die Straße muss jedoch nicht unbedingt nass sein, weil es geregnet hat. Trotzdem bieten sich diese erkannten Abhängigkeiten als neue Informationsquellen an. Diese kann man dazu nutzen, um durch z.B. *Interpolation/Extrapolation* die Lücken in den Daten zu schließen. Mit *Integration* kann man die unterschiedlichen Typen oder Quellen miteinander vereinen, was in dem Fall von Zeit und Raum geschieht um weitere Schlussfolgerungen zu erhalten.

2.3.2 Korrelationskoeffizient

Um bestimmte Abhängigkeiten zwischen Zeit und Ort herauszufinden, wird in Malik u. a. (2012) die Analyse mittels des Korrelationskoeffizient (*Pearson product-moment correlation coefficient*) vorgestellt. Die untersuchten Datensätze kommen jeweils aus den Bereichen Kriminalität, Verkehr und Zivildaten, wobei das ausgewählte Szenario in der Domäne der Strafverfolgung liegt. Das vorgestellte System kann also auf viele Arten von Daten (mit mehreren Variablen) aufsetzen.

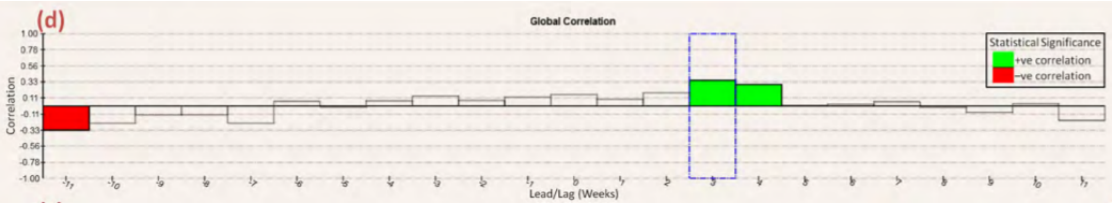


Abbildung 2.5: Untersuchung von Zusammenhängen in Kriminaldaten

Es wird dabei betont, dass es eine Herausforderung ist die potenziellen Beziehungen zu finden, weil einerseits die Daten unzureichend sind und andererseits die Komplexität des Problems zu hoch ist. Der Korrelationskoeffizient wird durch die folgende Formel berechnet:

$r_{xy} = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{(N-1)S_xS_y}$ Die Variable N ist die Länge der Zeitreihe und $\bar{x}$ oder $\bar{y}$ der Durchschnitt der jeweiligen Werte von x oder y. Das jeweilige S spiegelt die statistische Standardabweichung in der Region, wo x und y sich überschneiden wider. Damit man unterschiedliche zeitliche Datensätze untereinander vergleichbar macht, kann man zusätzlich einen Wert für die zeitliche Verschiebung nach vorne oder hinten angeben. Diesen sieht man z.B. in Abbildung 2.5 unter der x-Achse. Das nach der Formel berechnete Ergebnis hat eine große Aussagekraft, da es nur drei Aussagen geben kann. Einmal das die Daten sehr wahrscheinlich im Zusammenhang-stehen ($r_{xy} = +1$), eher nicht ($r_{xy} = 0$) oder ein gegenläufiger Zusammenhang ($r_{xy} = -1$) besteht. Wichtig für die Berechnung ist, dass die Zeitreihen die gleiche Länge aufweisen, was nicht immer der Fall ist. Dieses Problem umgehen die Autoren dadurch, dass sie z.B einfach die fehlenden Teile am Anfang mit den Daten von hinten auffüllen oder das sich nicht die ganzen Zeitreihen angeschaut werden, sondern nur die gerade überlappenden bzw. im Fokus stehenden gleichlangen Perioden. Um den Benutzer bei seiner Aufgabe zu unterstützen gibt es noch eine Gegenprüfung, ob die berechneten Werte unter der Verschiebung der Zeitreihe signifikant sind. Zudem kommt noch hinzu, dass man die Regionen der beiden untersuchten Zeitreihen zur Exploration angezeigt bekommt. Dies kann den Analysten zu neuen Theorien leiten, die vorher nicht einbezogen wurden. Das sieht man z.B. in Abbildung 2.6, wo der Bereich mit den meisten Einbrüchen in einem Gebiet mit vielen Drogendelikten liegt. Es ist nicht abwegig zu sagen, dass sich Drogenabhängige irgendwie Geld für ihren Drogenkonsum beschaffen müssen.

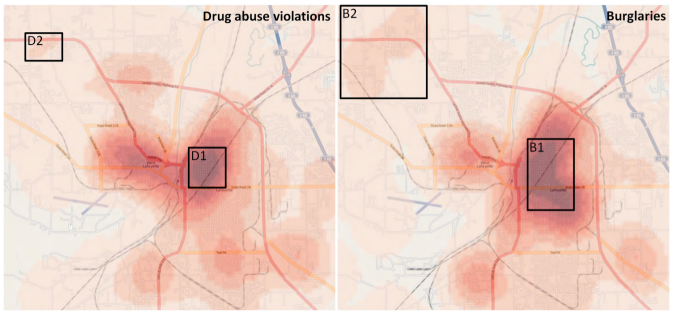


Abbildung 2.6: Vergleich zwischen Drogendelikten und Einbrüchen für das Jahr 2011

Die in Unterabschnitt 2.2.2 vorgestellte Methode wird in dieser Ausarbeitung nicht verwendet, sondern die eher gröbere Darstellung von Regionen durch *Heatmaps* (siehe Abbildung 2.6) mit Hilfe einer Technik zur Dichte-Schätzung von Andrienko u. a. (2010). Auch hier kann man mit Hilfe des Systems einfach Bereiche und die darin passierten Delikte miteinander auswählen und vergleichen. Zudem kann die zeitliche Skalierung auf Tage, Wochen oder Monate eingestellt werden, damit man auch aggregierte Datensätze untersuchen kann. Um den Einsatzzweck zu

verdeutlichen werden in Malik u. a. (2012) mehrere Fallstudien mit allen möglichen Erkenntnissen präsentiert, die man durch die Analyse von zeitlichen und räumlichen Daten erlangt.

2.3.3 Bewertung

Die größte Schwäche die bei dem Werkzeug sofort auffällt, ist die mögliche falsche Interpretation von Zusammenhängen die eigentlich nicht existieren, jedoch durch den Benutzer erzeugt wurden. Dem Benutzer ist es frei überlassen die Zeitreihen unter einander zu verschieben und anzupassen. Die Analyse bedarf also einer Person mit großer Erfahrung in dem jeweils zu untersuchenden Gebiet. Ebenso braucht man wie in Unterabschnitt 2.2.3 einen Austausch mit Domänenexperten zur besseren Implementierung eines solchen *Visual Analystics* Systems. Für die Evaluierung der eigenen Arbeit reicht es nicht nur schlüssige Fallstudien vorzuzeigen, diese sollten auch untermauert sein. Im Gegensatz dazu kann man jedoch solche Ansätze gut für die Exploration von vielen Daten nutzen.

3 Fazit

Im ersten Konzept in Unterabschnitt 2.1.2 wurden eindimensionale zeitliche Daten ohne fehlende Messwerte untersucht, um daraus Vorhersagen über zukünftige Entwicklungen zu bekommen. Die verwendete Zeitreihenanalyse und die Verbindung dieser mit *Visual Analytics* ist sehr gut dokumentiert und fachgerecht beurteilt durch die Hilfe von Experten aus dem Gebiet der Statistik. Das beschriebene Vorgehen lässt sich gut weiterverwenden, um eigenen Ergebnisse zu evaluieren. Die α-shapes in Unterabschnitt 2.2.2 wurden für die Suche von interessanten Mustern bzw. Regionen eingesetzt. Das Resultat sah visuell ansprechend aus und verfälschte auch nicht die dahinter liegenden Informationen, da die Lücken in den Daten sichtbar sind. In dem dritten Abschnitt wird die Zeitreihenanalysen mit der Suche von periodischen Mustern verbunden, um den Zusammenhang zwischen dem Ort und der Zeit zu analysieren. Dabei enthalten auch hier die Daten Ausreißer oder möglicherweise auch Fehler. Bei den α-shapes wurde die Skalierung über die farblich hervorgehobene Hierarchie dargestellt, die der Benutzer frei einstellen konnte. Zu der räumlichen Skalierung kommt in Unterabschnitt 2.3.2 die Möglichkeit bestimmte Perioden und Orte auszuwählen hinzu. Dadurch kann man sich frei in den Daten explorativ bewegen, um vielleicht so auf unbekannte Abhängigkeiten zu stoßen. Diese Aufgabe muss jedoch von einem erfahren Analysten oder Domänenexperten gemacht werden, was bei eigenen Projekten zu beachten ist. Die richtige Bewertung der erbrachten Arbeit könnte auch so realisiert werden, dass man vorher präparierte Datensätze von Testpersonen analysieren lässt und den Fortschritt dabei beobachtet und bewertet. Jedenfalls stellen nicht die Methoden die schwerste Hürde dar, sondern der Nachweis der Richtigkeit des Systems und der gemachten Schlussfolgerungen.

Literaturverzeichnis

[Andrienko u. a. 2010] Andrienko, Gennady ; Andrienko, Natalia ; Bremm, Sebastian ; Schreck, Tobias ; Von Landesberger, Tatiana ; Bak, Peter ; Keim, Daniel: Space-in-Time and Time-in-Space Self-Organizing Maps for Exploring Spatiotemporal Patterns. In: *Computer Graphics Forum* Bd. 29 Wiley Online Library, 2010, S. 913–922 9

[Bogl u. a. 2013] Bogl, M. ; Aigner, W. ; Filzmoser, P. ; Lammarsch, T. ; Miksch, S. ; Rind, A.: Visual Analytics for Model Selection in Time Series Analysis. In: *Visualization and Computer Graphics, IEEE Transactions on* 19 (2013), Dec, Nr. 12, S. 2237–2246. http://dx.doi.org/10.1109/TVCG.2013.222. – DOI 10.1109/TVCG.2013.222. – ISSN 1077–2626 2, 4

[Box u. a. 2013] Box, George E. ; Jenkins, Gwilym M. ; Reinsel, Gregory C.: *Time series analysis: forecasting and control*. John Wiley & Sons, 2013 3

[Keim u. a. 2010] Keim, Daniel A. ; Kohlhammer, Jörn ; Ellis, Geoffrey ; Mansmann, Florian: *Mastering The Information Age-Solving Problems with Visual Analytics*. Florian Mansmann, 2010 2, 5, 8

[Krasnoshchekov u. Polishchuk 2008] Krasnoshchekov, Dmitry ; Polishchuk, Valentin: Robust curve reconstruction with k-order α-shapes. In: *Shape Modeling and Applications, 2008. SMI 2008. IEEE International Conference on* IEEE, 2008, S. 279–280 6

[Malik u. a. 2012] Malik, A. ; Maciejewski, R. ; Elmqvist, N. ; Jang, Yun ; Ebert, D.S. ; Huang, W.: A correlative analysis process in a visual analytics environment. In: *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, 2012, S. 33–42 8, 10

[Packer u. a. 2013] Packer, E. ; Bak, P. ; Nikkila, M. ; Polishchuk, V. ; Ship, H.J.: Visual Analytics for Spatial Clustering: Using a Heuristic Approach for Guided Exploration. In: *Visualization and Computer Graphics, IEEE Transactions on* 19 (2013), Dec, Nr. 12, S. 2179–2188. http://dx.doi.org/10.1109/TVCG.2013.224. – DOI 10.1109/TVCG.2013.224. – ISSN 1077–2626 5