Projektdokumentation im Modul Semantic Web

Einflüsse auf die Kriminalität in den Counties von England

Anne Matthes

10.07.2015

Recherchefragestellung: Welche sozialen und lokalen Kriterien nehmen Einfluss auf die Häufigkeit von Kriminalitätsvorfällen in britischen Regionen? Es ist zu überprüfen, ob eine Abhängigkeit zur Arbeitslosigkeit und Bevölkerungszahl besteht.
(Fokus: Counties von England)

1 Inhaltliche Interpretation der Fragestellung

Im Rahmen dieser Arbeit soll untersucht werden, ob Abhängigkeiten der Häufigkeit von kriminellen Vorkommen zu Kennzahlen, wie die Bevölkerungszahl und Arbeitslosigkeit, bestehen. Zur Untersuchung werden die Counties von England in Betracht gezogen. Für diese sind Datenquellen für die drei Kriterien *Kriminalitätsvorkommen*, *Bevölkerungszahl* und *Arbeitslosigkeit* im Zeitraum vom Dezember 2010 bis Dezember 2014 lückenlos auf Monatsbasis vorhanden. Eine mögliche Erwartungshaltung ist, dass mit zunehmender Arbeitslosigkeit und Bevölkerungszahl auch die Kriminalität steigen kann.
Weitere denkbare Ansätze wären die Analyse des zeitlichen Verlaufs der drei Kennzahlen oder die Untersuchung der Kriterien zu verschiedenen Jahreszeiten. Auch können Vergleiche der einzelnen Counties zu Gesamtengland oder Großbritannien aufgestellt oder weitere Datenquellen für Wales, Schottland und Nordirland hinzugenommen werden. Diese Ansätze wurden im Umfang dieses Projekts nicht ausgearbeitet, da diese den zeitlichen und inhaltlichen Rahmen der Projektarbeit übersteigern würden.

2 Relevante Datenquellen

An dieser Stell erfolgt eine Auflistung aller relevanten Datenquellen und deren Beschreibung.

2.1 Daten zum Kriminalitätsvorkommen in Großbritannien

Die Datenbank von *DATA.POLICE.UK* stellt aktuell sämtlichen Straftaten vom Dezember 2010 bis April 2015 von England, Wales, Schottland und Nordirland zur Verfügung. Über ein Formular lässt sich auswählen, welche Daten gewünscht sind (Orte und Zeitraum). Anschließend lässt sich ein ZIP-Ordner downloaden, welcher in einer komplexen Ordnerstruktur sämtliche CSV-Dateien (pro Ort und Monat) enthält. **?**

Link	`http://data.police.uk/data/`
Datenformat	CSV (in komplexer Ordnerstruktur)
Schnittstelle	HTTP (früher API)
Lizenz	Open Government Licence
Open Data	⋆⋆⋆

2.2 Daten zur Bevölkerung in Großbritannien - Censusdaten

Censusdaten von Großbritannien werden durch die Nomis-Statistiken bereitgestellt. Über den Menüpunkt *Data downloads ¿ Query* können ähnlich wie bei der Datenquelle der Kriminalitätsvorkommen über ein Formular die gewünschten Daten angepasst werden. Dazu ist zunächst das Dataset *Population Estimates ¿ mid-year population estimates* zu wählen. Es stehen die Daten von 1981 bis 2014 zur Verfügung. Die Daten können als XLSX, CSV, im Web Browser, als Map, TSV oder über die Nomis API abgerufen werden. **?**

Link	`https://www.nomisweb.co.uk/`
Datenformat	XLSX, CSV, TSV, HTML
Schnittstelle	HTTP, API
Lizenz	Open Government Licence
Open Data	⋆⋆⋆⋆

2.3 Daten zur Arbeitslosigkeit in Großbritannien - London Datastore

Daten zu Arbeitslosigkeit in England lassen sich vom *London Datastore* abrufen. Unter dem Menüpunkt *Data* lässt sich das passende Dataset unter *Employment and Skills ¿ Unemployment Rate, Region* zum Download als XLS finden. Für die einzelnen Regionen Englands sind Zahlen der Arbeitslosigkeit monatsweise im Zeitraum von April 2009 bis April 2015 zu finden. **?**

Link	`http://data.london.gov.uk/dataset/`
Datenformat	XLS
Schnittstelle	HTTP
Lizenz	Open Government Licence
Open Data	⋆⋆

2.4 Daten zur Zuordnung der Counties von England zu den Regionen - Wikipedia

Im Onlinelexikon *Wikipedia* lässt sich eine vollständige Liste der Regionen Englands mit ihren zugehörigen Counties finden. Aus dieser HTML-Seite können die Daten geparst und die nötigen Informationen für das Projekt gewonnen werden. **?**

Link	`https://de.wikipedia.org/wiki/Verwaltungsgliederung_Englands/`
Datenformat	XLS
Schnittstelle	HTTP
Lizenz	GNU Free Documentation License
Open Data	⋆⋆⋆⋆

3 Extraktion relevanter Daten und Import in einen Triplestore

Die Extraktion und Aufbereitung der Daten für den Triplestore *Apache Fuseki* **?** erfolgt mittels verschiedener Methoden einer für das Projekt entwickelten Java Applikation.

3.1 Extraktion der Daten von Kriminalitätsvorkommen

Zur Extraktion der Kriminalitätsdaten müssen viele einzelne CSV-Dateien innerhalb einer komplexen Ordnerstruktur ausgelesen werden. Mithilfe der Methode **listFiles()**

kann durch die Verzeichnisse navigiert und anschließend die CSV-Dateien mittel der *Java Excel API* **?** ausgewertet werden.

Anhand der Bezeichnung der CSV-Files lässt sich das Countie und der betreffende Monat bestimmen.

```
// Ausschnitt aus der Klasse LoadCrimeData()
public ArrayList<String[]> loadData(String res) {

      File d = new File(res);
      File[] dirArray = d.listFiles();

      for (File dir : dirArray) {
         File f = new File(dir.toString());
         File[] fileArray = f.listFiles();

         for (File file : fileArray) {
            String[] DataArray = new String[3];

            DataArray[0] = getLocation(file);
            DataArray[1] = getDate(file);
            DataArray[2] = getCrimeNumber(file);

            CrimeData.add(DataArray);
         }
      }

      return CrimeData;
   }
```

Die Methode **getLocation()** liest u. a. die County-Bezeichnungen ein und passt sie zur späteren Verlinkung an. Es werden Leerzeichen und Unterstriche ersetzt und Anfangsbuchstaben der Teilbezeichnungen groß geschrieben.
Die Anzahl der Kriminalitätsvorkommen wird durch die Methode **getCrimeNumer()** ermittelt. Hier wird die Anzahl der Einträge in den CSV-Files mittels der *Java Excel API* ermittelt.
Anschließend werden die Werte *Location*, *Date* und *CrimeNumber* in ein **StringArray** gespeichert und dieses zur weiteren Verarbeitung an eine **ArrayList** übergeben. [1]

[1] vgl. https://github.com/annefresa/semanticweb/blob/master/src/LoadDataSets/LoadCrimeData.java

3.2 Extraktion der Bevölkerungsdaten

Bei der Extraktion der Bevölkerungszahlen wird ebenfalls die *Java Excel API* verwendet. Ausgelesen werden in diesem Fall wieder die Werte für *Location* und *Date* sowie *PopulationNumer* und **LocationType**.
Alle Datensätze mit dem **LocationType** *üacounty"* werden ebenfalls wieder in eine **ArrayList** aufgenommen. [2]

```
// Ausschnitt aus der Klasse LoadPopulationData() und Methode
// public ArrayList<String[]> loadData(String res)
String[] DataArray = new String[4];

   DataArray[0] = formatLocation(Location);
   DataArray[1] = Date;
   DataArray[2] = formatNumbers(PopulationNumber);
   DataArray[3] = getLocationType(sheet.getCell(0, i)
              .getContents().split(":")[0].substring(0, 3));

   if (DataArray[3].equals("uacounty"))
      PopulationData.add(DataArray);
//...
```

3.3 Extraktion der Daten der Arbeitslosigkeit

Die Daten der Arbeitslosigkeit werden analog wie die Bevölkerungsdaten erfasst.

```
// Ausschnitt aus der Klasse LoadUnemploymentData() und Methode
// public ArrayList<String[]> loadData(String res)
String[] DataArray = new String[3];

   DataArray[0] = formatRegion(Region);
   DataArray[1] = convertDate(Date);
   DataArray[2] = formatNumbers(UnemploymentNumber);

   UnemploymentData.add(DataArray);
//...
```

Dabei werden die Methoden **formatRegion**, **convertDate** und **formatNumbers** verwendet. Somit werden Leerzeichen bei Regionsbezeichnungen durch Unterstriche ersetzt, das Datum für die Verlinkungen angepasst und für spätere Auswertungen Kommas aus

[2] vgl. https://github.com/annefresa/semanticweb/blob/master/src/LoadDataSets/LoadPopulationData.java

den noch als **String** abgespeicherten Arbeitslosenzahlen entfernt. [3]

3.4 Extraktion der Daten mit den Counties-Regionen-Beziehungen

Um die Daten der Counties-Regionen-Beziehungen zu erfassen, muss der HTML-Code von Wikipedia geparst werden. Dazu wird das Framework *JSOUP* verwendet.

```
// Ausschnitt aus der Klasse LoadLocationRelationData()
public ArrayList<String[]> loadData(String[] res) {

      try {
         Document doc = Jsoup.connect(res[0]).get();
         splittedHtml = doc.body().toString().split("\n");

         for (...)
         //...

            if (splittedHtml[i].contains("<h3>")) {
               Region = formRegionName(splittedHtml[i].substring(38,
                     splittedHtml[i].indexOf("\">")));
            }

            if (splittedHtml[i].contains("<li>")) {
               County = formCountyName(splittedHtml[i].substring(
                     splittedHtml[i].indexOf("\">") + 2,
                     splittedHtml[i].indexOf("</a>")));

               String[] DataArray = new String[2];

               DataArray[0] = County;
               DataArray[1] = Region;

               LocationData.add(DataArray);
            }
         }

      } catch (IOException e) {
         e.printStackTrace();
      }
      //...
```

Auch hier werden die relevanten Daten zuerst in einem **StringArray** zusammengefasst und anschließend in einer **ArrayList** gespeichert. Mittels der Methoden **formRegion-**

[3] vgl. https://github.com/annefresa/semanticweb/blob/master/src/LoadDataSets/LoadUnemploymentData.java

Name() und **formCountyName()** werden auch in dieser Klasse die Bezeichnungen für das folgende Matching angepasst. [4]

4 Verlinkung von Ressourcen

Um die Daten später in den Triplestore *Apache Fuseki* laden zu können, wurde die Daten zunächst in der Klasse **OntologyBuilder()** im RDF-Schema in eine Datei geschrieben. Folgender Code zeigt verkürzt das Prinzip der Erstellung eines RDF-Files mithilfe des hier verwendeten *Jena Frameworks* **?**:

```
// Ausschnitt aus der Klasse OntologyBuilder()
      String uri = "http://www.imn.htwk-leipizg.de/~amatthes/semweb/schema#";

      OntModel model = ModelFactory
            .createOntologyModel(OntModelSpec.OWL_DL_MEM_RULE_INF);
      model.setNsPrefix("am", uri);

      Resource crimeRes = model.createResource(uri + "Crime");
      model.add(crimeRes, RDF.type, RDFS.Class);
      //...

      Property hasCrimeCounty = model.createProperty(uri + "hasCrimeCounty");
      model.add(hasCrimeCounty, RDF.type, RDF.Property);
      model.add(hasCrimeCounty, RDFS.domain, crimeRes);
      model.add(hasCrimeCounty, RDFS.range, countyRes);
      //...

      for (String[] entity : crimeData) {
         Resource crime = model.createResource(uri + "crime_" + entity[0]
               + "_" + entity[1]);
         Resource county = model.createResource(uri + entity[0]);
         Resource date = model.createResource(uri + entity[1]);
         model.add(crime, hasCrimeCounty, county);
         model.add(crime, hasCrimeDate, date);
         model.add(crime, hasCrimeAmount,
               model.createTypedLiteral(Double.parseDouble(entity[2])));
         model.add(crime, RDF.type, crimeRes);
      }
      //...
      try {
         model.write(new FileOutputStream(
               "/Users/anmt/Desktop/semwebOntology.rdf"));
      //...
```

[4] vgl. https://github.com/annefresa/semanticweb/blob/master/src/LoadDataSets/LoadLocationRelationData.java

Die drei für das Ergebnis relevanten Klassen *Crime*, *Population* und *Unemployment* werden dabei über die Eigenschaften bzw. Klassen *Date* und *County* verknüpft.
Da jedoch der Arbeitslosigkeit keine Counties, sondern nur Regions zugeordnet werden können, muss noch eine Klasse *Region* erstellt werden. Über diese kann anschließend eine Verlinkung der Arbeitslosenzahlen mit den anderen Kennzahlen durchgeführt werden.

Crime	hasCrimeCounty	County
Population	hasPopCounty	County
Unemployment	unempInRegion	Region
County	countyInRegion	Region
Crime	hasCrimeDate	Date
Population	hasPopDate	Date
Unemployment	hasUnempDate	Date

Die resultierende RDF-Datei ist auf Github zu finden. [5]

5 Anfrage an die Forschungswissensbasis

5.1 SPARQL-Anfrage

Um die Daten für eine grafische Darstellung aufbereiten zu können, sind möglichst viele Wertepaare von Kriminalitäts- und Bevölkerungszahlen oder Kriminalitäts- und Arbeitslosenzahlen zu finden. Diese werden jeweils über den Ort und das Datum gemacht:

Kriminalität - Bevölkerung (1029 Wertepaare)

```
PREFIX am: <http://www.imn.htwk-leipizg.de/~amatthes/semweb/schema#>

SELECT ?CrimeAmount ?PopAmount
WHERE{
  ?crime am:hasCrimeDate ?Date .
  ?pop am:hasPopDate ?Date .

  ?crime am:hasCrimeCounty ?County .
  ?pop am:hasPopCounty ?County .

  ?crime am:hasCrimeAmount ?CrimeAmount .
  ?pop am:hasPopAmount ?PopAmount .
}
```

[5] vgl. https://github.com/annefresa/semanticweb/blob/master/rdf/semwebOntology.rdf

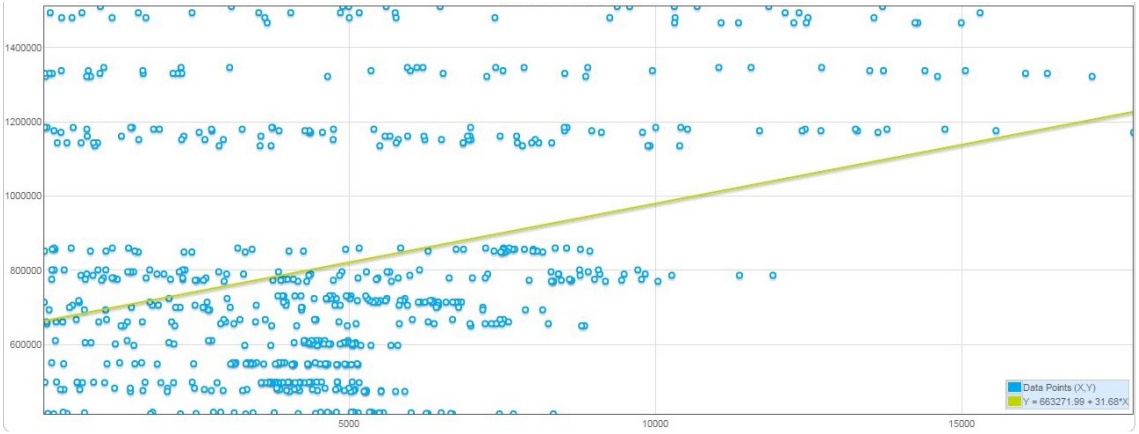
Kriminalität - Arbeitslosigkeit (882 Wertepaare)

```
PREFIX am: <http://www.imn.htwk-leipizg.de/~amatthes/semweb/schema#>

SELECT ?CrimeAmount ?UnempAmount
WHERE{
  ?crime am:hasCrimeDate ?Date .
  ?unemp am:hasUnempDate ?Date .
  ?crime am:hasCrimeCounty ?County .

  ?unemp am:unempInRegion ?Region .
  ?County am:countyInRegion ?Region .

  ?crime am:hasCrimeAmount ?CrimeAmount .
  ?unemp am:hasUnempAmount ?UnempAmount .
}
```

Auf Github können diese und noch weitere Beispielanfragen gefunden werden. So lassen sich die einzelnen Zahlen z. B. auch auf die Regionen zusammenfassen oder im zeitlichen Verlauf darstellen. Noch zahlreiche weitere Abfragen sind denkbar. [6]

5.2 Ergebnis der Anfrage

Kriminalität - Bevölkerung

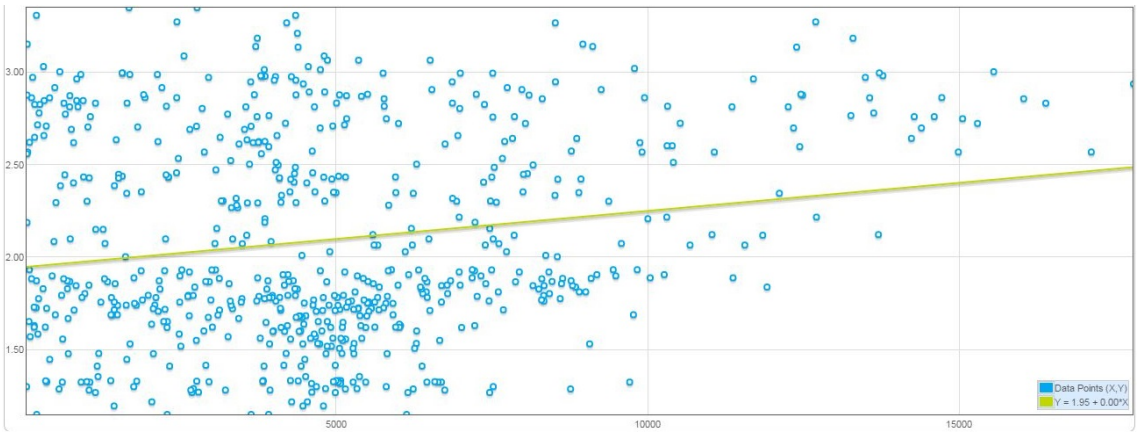


An der x-Achse lassen sich die Werte der Kriminalitätsvorkommen und an der y-Achse die Bevölkerungszahlen ablesen. Jeder eingezeichnete Punkt stellt ein Wertepaar dar.

[6] vgl. `https://github.com/annefresa/semanticweb/tree/master/rdf`

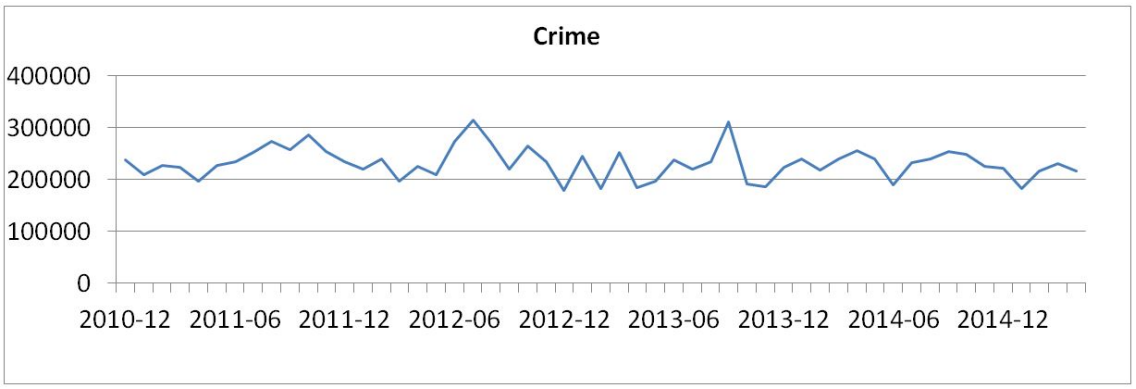
Die Grafik wurde mit der Anwendung *Calculate Linear Regression Graph Scatter Plot* erstellt und die grüne Geradenfunktion automatisch eingefügt. [7]

Kriminalität - Arbeitslosigkeit



Die Kriminalitätsdaten sind auch in dieser Grafik an der x-Achse und die Arbeitslosenzahlen an der y-Achse abzulesen.

Zeitlicher Verlauf der Kriminalitätsvorkommen



Dieses Diagramm veranschaulicht den zeitlichen Verlauf der Kriminalitätsvorkommen (mit Microsoft Excel erstellt). Die zugehörigen Daten sind in einer CSV-Datei auf Github zu finden. [8]

[7] vgl. `http://www.livephysics.com/tools/mathematical-tools/calculate-linear-regression\ \-graph-scatter-plot-line-fit/`

[8] vgl. `https://github.com/annefresa/semanticweb/blob/master/results/Daten_zeitlicher_Verlauf.csv/`

6 Interpretation und Zusammenfassung

Kriminalität - Bevölkerung
Eine eindeutige Abhängigkeit zwischen der Kriminalität und Bevölkerungszahl ist nicht zu erkennen. Es ist jedoch zu erkennen, dass nur Counties mit einer sehr hohen Bevölkerung auch eine extrem hohe Kriminalität vorweisen. Somit lässt sich wage die Behauptung aufstellen, dass die Kriminalität mit zunehmender Bevölkerung ebenfalls zunehmen kann. Die Separierung der Ergebnisse in zwei Teilmengen beruht zum einen auf die vorhandene Bevölkerungsgröße der Counties. Es existieren in diesem Fall keine Städte mit mittlerer Bevölkerungszahl und verursachen somit die Lücke im Graphen. Die zeilenweise Anordnung entsteht, da sich über den Zeitraum von vier Jahren die Bevölkerungszahl pro Ort nicht stark ändert, sondern eher schwankend ist.

Kriminalität - Arbeitslosigkeit
Die Abhängigkeiten zwischen Kriminalität und Arbeitslosigkeit ist sehr weit gestreut. Dennoch kann die Tendenz erkannt werden, dass eine erhöhte Arbeitslosigkeit tatsächlich die Kriminalität steigern kann.

Zeitlicher Verlauf der Kriminalitätsvorkommen
Die zeitliche Gegenüberstellung der Kriminalitätsvorkommen lässt keine eindeutigen Schlussfolgerungen zu. Es gibt keine periodischen Wiederholungen, die sich auf den Monat oder die Jahreszeit beziehen. Generell können Schwankungen durch viele Faktoren beeinflusst werden, wie z. B. neue Gesetze oder die wirtschaftliche Lage Englands.

Weitere Auswertungen
Die Stadt mit der geringsten Kriminalität ist die *City of London*. Allerdings wurden kein Bezug zur Bevölkerungszahl genommen. [9]
Dem entgegen steht *West Yorkshire* mit dem höchsten Kriminalitätsvorkommen. [10]
Die meisten Kriminalitätsfälle gab es im September 2013 im County *Metropolitan* mit **37294** Vorfällen. [11]
In *Northamptonshire* wurden im Juli 2014 hingegen gerade einmal **32** kriminelle Taten erfasst.

Fazit
Mithilfe der aufbereiteten Daten können viele verschiedene und umfangreiche Abfragen und Analysen durchgeführt werden. Jedoch ist zu beachten, dass Unvollständigkeiten beim Matching durch verschiedene Bezeichnungen entstehen können. **?**

[9] vgl. `https://github.com/annefresa/semanticweb/blob/master/sparql_requests/crime-safest-counties.rq`
[10] vgl. `https://github.com/annefresa/semanticweb/blob/master/sparql_requests/crime-unsafest-counties.rq`
[11] vgl. `https://github.com/annefresa/semanticweb/blob/master/sparql_requests/crime-date-ranking.rq`