Karten (und andere Daten) aus OpenStreetMap in SAS nutzen – am Beispiel von Postleitzahlgebieten

Thorben Breitkreuz
AQUA - Institut für angewandte Qualitätsförderung und Forschung im Gesundheitswesen GmbH
Maschmühlenweg 8-10
37073 Göttingen
t.breitkreuz@aqua-institut.de

Thomas Grobe
AQUA - Institut für angewandte Qualitätsförderung und Forschung im Gesundheitswesen GmbH
Maschmühlenweg 8-10
37073 Göttingen
t.grobe @aqua-institut.de

Riccardo Gezzi
AQUA - Institut für angewandte Qualitätsförderung und Forschung im Gesundheitswesen GmbH
Maschmühlenweg 8-10
37073 Göttingen

Susanne Steinmann
AQUA - Institut für angewandte Qualitätsförderung und Forschung im Gesundheitswesen GmbH
Maschmühlenweg 8-10
37073 Göttingen
s.steinmann @aqua-institut.de

Zusammenfassung

Der vorliegende Beitrag zeigt die direkte Extraktion und Aufbereitung von Informationen aus OSM-Daten in SAS am Beispiel der Postleitzahlregionen Deutschlands. Damit wollen die Autoren exemplarisch einen auch vielfältig anderweitig nutzbaren Zugang zu OSM-Daten darstellen.

Schlüsselwörter: SAS/GRAPH, PROC GMAP, SAS Map Data Set, OpenStreetMap, XML, Postleitzahlen, Karten, Choropletengrafik

1 Motivation

Die geografische Darstellung von Forschungsergebnissen ist ein probates Mittel, um Unterschiede des Untersuchungsgegenstandes in regionalen Strukturen auf anschauliche Weise aufzuzeigen. Häufig verwendet werden dabei sogenannte Choropletengrafiken. Choropletengrafiken verwenden Farbabstufungen und/oder Muster, um die – zumeist kategorisierten – Ergebnisse einer Untersuchung auf einer Karte darzustellen. Dabei werden Kartengrundlagen verwendet, die eine Abgrenzungen von Gebieten, z.B. von Bundesländern oder Kreisen, ermöglichen. In Abbildung 1 findet sich ein Beispiel einer solchen Darstellung auf der Ebene von Kreisen für relative Abweichung der Verordnungsrate einer bestimmten Arzneimittelgruppe im Vergleich zur erwarteten Rate. Regionale Unterschiede sind dabei durch die Farbabstufungen (Grauabstufungen) leicht erkennbar.

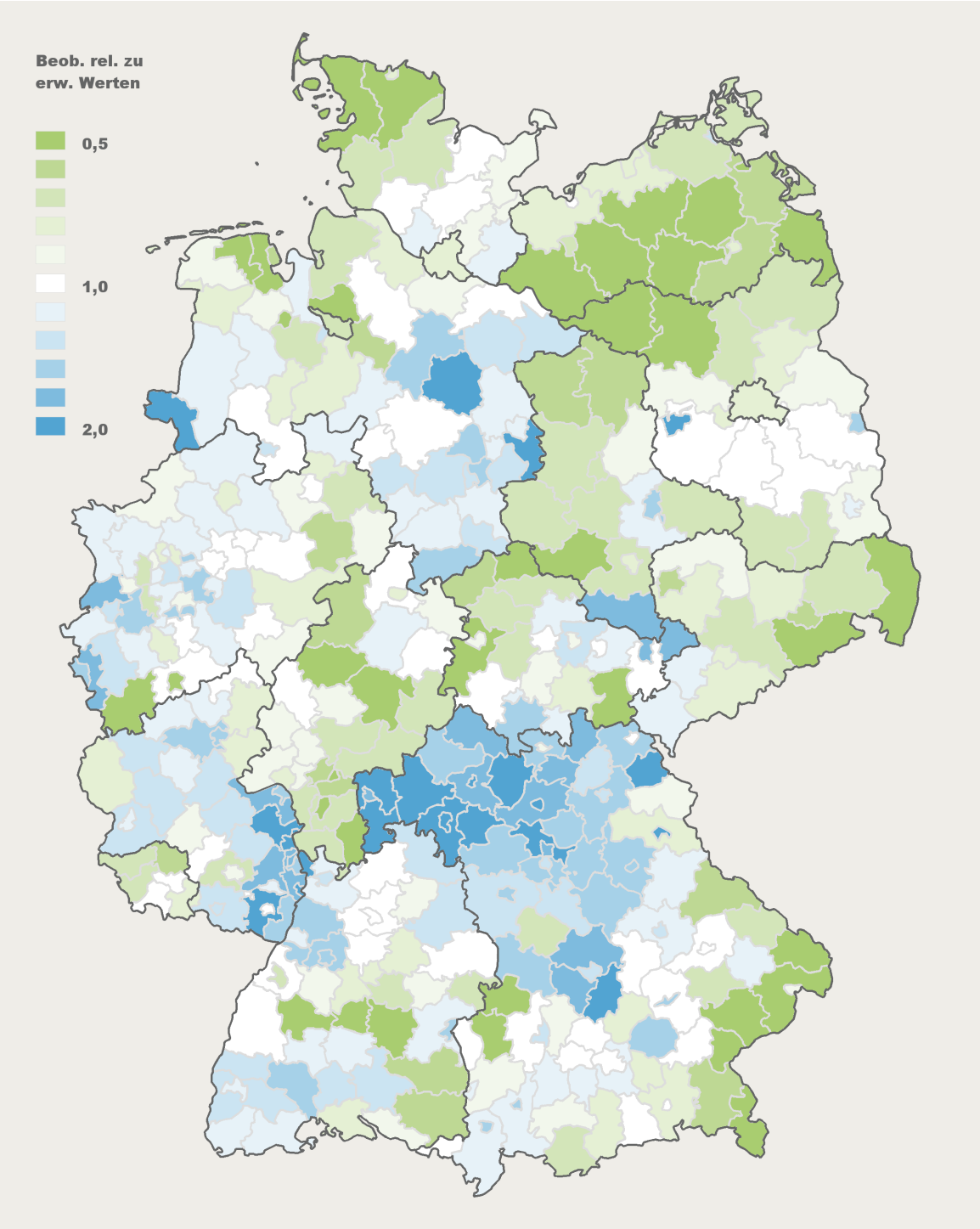

Abbildung 1: Relative Abweichungen beobachteter Methylphenidat-Verordnungsraten von erwarteten Raten in Kreisen 2014 (Altersgruppen 0 bis 19 Jahre, indirekt standardisiert)[1]

SAS liefert eine Reihe von Karten, die als Grundlage solcher Darstellungen dienen können. Auf diese Karten – es handelt sich dabei um zwei normale Data Sets, die lediglich eine bestimmte Struktur aufweisen – kann über die Maps-Library zugegriffen werden,

[1] Quelle: BARMER GEK Arztreport 2016 - Schwerpunkt: Alter und Schmerz. Schriftenreihe zur Gesundheitsanalyse, Band 37, BARMER GEK, 2016, Berlin.

die Teil der Standardinstallation von SAS ist. So wird für Deutschland beispielsweise eine Karte mitgeliefert, die bis hinunter auf Kreisebene unterteilt ist (vgl. Abbildung 1). Die Darstellung der Karten kann dabei mit der im SAS/GRAPH-Modul enthaltenen Prozedur GMAP erfolgen. Was tut man aber, wenn man eine andere regionale Unterteilung, wie z.B. die auf Postleitzahlenebene, benötigt?

2 OpenStreetMap als Quelle geografischer Daten

Mit OpenStreetMap (OSM) stehen als Datengrundlage für SAS-Karten umfangreiche geografisch zugeordnete Informationen aus allen Regionen der Welt zur Verfügung, die von Jedermann genutzt werden können. Sämtliche Daten zu OSM werden (auch) in einem XML-Format im Internet zum Download bereitgestellt. Da SAS in der Lage ist XML-Dateien zu laden, lag der Versuch nahe, OpenStreetMap-Daten als Grundlage für eigene Map Data Sets in SAS zu verwenden.
OSM-Daten beinhalten im Wesentlichen drei Grundtypen von Elementen: 1. Knoten (nodes), die jeweils einzelne geokodierte Punkte auf der Erdoberfläche definieren, 2. Wege (ways), die bestimmte Abfolgen von Knoten definieren (z.B. um Straßen- und Grenzabschnitte oder Gebäudeumrisse darzustellen), 3. Relationen (relations), die dazu dienen, Zusammenhänge von bestimmten Wegen, aber auch Knoten und andere Relationen zu beschreiben, um größere Einheiten aus vielen Teilen zusammensetzen zu können (z.B. Ländergrenzen). Alle drei Elemente besitzen typenbezogen eindeutige Identifier und können jeweils durch Tags und Attribute weiter spezifiziert und beschrieben werden (z.B. eine Relation als die Grenze eines bestimmten Landes).

3 Abbildung von Postleitzahlengebieten in OpenStreetMap

Um eine „Übersetzung“ der OSM-Datenstruktur in ein SAS Map Data Set zu erreichen, kommt man allerdings nicht umhin, sich genauer mit den sogenannten „Map Features“[2] von OSM zu beschäftigen, welche die genauere Struktur von einzelnen Elementen auf der Karte beschreiben. Im vorliegenden Fall müssen also Elemente, die für die Darstellung von Postleitzahlengebieten im Gebiet von Deutschland notwendig sind, identifiziert werden.
Auf der genannten Webseite findet man die Information, dass Postleitzahlengebiete als Relationen vom Typ Grenze (boundary) mit der Eigenschaft Postleitzahl (postal_code) gekennzeichnet sind. Jede Relation bildet dabei genau ein Postleitzahlengebiet ab. Die Umrisse des Gebietes werden durch Wege gebildet, die wiederum auf eine Abfolge von einzelnen Knoten verweisen, die benötigt werden, um diese Wege zu zeichnen.
Am Beispiel eines Postleitzahlengebietes (Postleitzahl 34119 der Stadt Kassel) soll dies im Folgenden näher erläutert werden. In Abbildung 2 ist das Gebiet der Postleitzahl 34119 Kassel zu sehen. Die orange Umrisslinie kennzeichnet dabei die zugeordnete Region.

[2] Für Deutschland: http://wiki.openstreetmap.org/wiki/DE:Map_Features

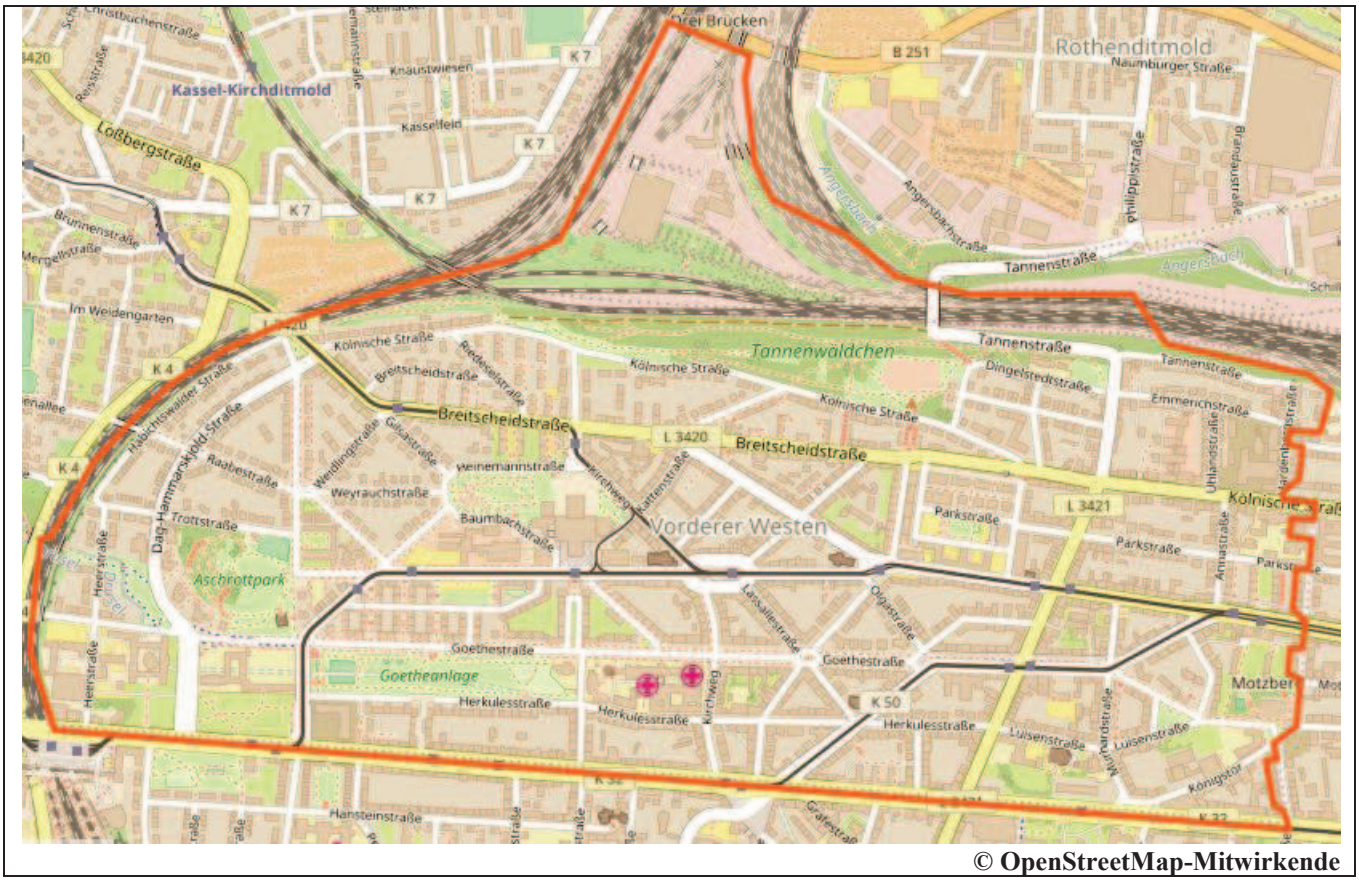

Abbildung 2: mit Postleitzahlgebiet 34119 der Stadt Kassel

Exportiert man die dazugehörigen Informationen als XML-Datei, mit der auf der Webseite (www.openstreetmap**2:** Kartenausschnitt.org/export) bereitgestellten Funktionalität, so ergibt dies die nachstehenden XML-Daten[3]:

```
<relation id="1189969" visible="true" version="13"
     changeset="43618320" timestamp="2016-11-13T23:43:47Z"
     user="wambacher" uid="201359">
  <member type="way" ref="79123010" role="outer"/>
  <member type="way" ref="453136782" role="outer"/>
  <member type="way" ref="79123005" role="outer"/>
  <member type="way" ref="79123008" role="outer"/>
  <member type="way" ref="79123023" role="outer"/>
  <member type="way" ref="79122997" role="outer"/>
  <member type="way" ref="79123006" role="outer"/>
  <tag k="boundary" v="postal_code"/>
  <tag k="note" v="34119 Kassel"/>
  <tag k="postal_code" v="34119"/>
  <tag k="postal_code_level" v="8"/>
  <tag k="type" v="boundary"/>
</relation>
```

Umschlossen von einem `<relation>…</relation>` Tag-Paar (mit Metadaten zum Erstellungszeitpunkt, Version und Benutzer, der dieses Objekt zuletzt geändert hat),

[3] Das oberste hierarchische Element vom Typ OSM wurde der Übersichtlichkeit halber entfernt.

werden mehrere `<member …/>` tags vom Typ Weg referenziert. Weiterhin sind einige `<tag …/>` Einträge mit Angaben zum Gebiet enthalten.

Da die Relation mehr als eine Referenz auf einen Weg enthält, bedeutet dies, dass das Postleitzahlengebiet 34119 Kassel nicht aus einer durchgehenden Umrisslinie besteht, sondern sich diese aus – in diesem Fall sieben – einzelnen Wegen zusammensetzt. In Abbildung 2 sind diese einzelnen Linien, durch schwarze Markierungen voneinander abgegrenzt und mit den Nummern der Wege (schwarze Schrift) versehen dargestellt.

Abbildung 3: Abschnitte der Umrisslinie des Postleitzahlgebietes 34119 Kassel

Der erste Weg (`79123010`) aus Abbildung 3 ist im nachfolgenden Abschnitt der XML-Daten dargestellt. Dieser enthält Verweise auf 13 Knoten (`<nd ref="..."/>`), die ihrerseits notwendig sind, um diesen Weg zu zeichnen.

```
<way id="79123010" visible="true" version="5" changeset="43618320"
timestamp="2016-11-13T23:43:46Z" user="wambacher" uid="201359">
  <nd ref="924504450"/>
  <nd ref="924504256"/>
  <nd ref="924504137"/>
  <nd ref="925551353"/>
  <nd ref="924504315"/>
  <nd ref="924504446"/>
  <nd ref="1899969160"/>
  <nd ref="925551290"/>
```

```
  <nd ref="924504248"/>
  <nd ref="1899969141"/>
  <nd ref="925551341"/>
  <nd ref="924504416"/>
  <nd ref="925551287"/>
  <tag k="boundary" v="postal_code"/>
 </way>
</osm>
```

Einer dieser Knoten (`924504137`) des ersten Weges (`79123010`) ist als Beispiel in der **Abbildung 3**Abbildung 3 mit einem gelben Kreis markiert und beschriftet.

Die XML-Entsprechung dieses Knotens enthält dann schließlich die Koordinaten als Längen- und Breitengradangabe (`lon` = longitude = Längengrad; `lat` = lattitude = Breitengrad) im Format des World Geodetic Systems 1984 (WGS 84).

```
<node id="924504137" visible="true" version="4" changeset="7575813"
timestamp="2011-03-16T13:41:43Z" user="BurnyB" uid="83876"
lat="51.3216733" lon="9.4603761"/>
```

Zu beachten ist, dass sich alle Elemente (Relationen, Wege, Knoten) an verschiedenen Stellen des XML-Dokuments befinden. Innerhalb einer Relation oder eines Weges ist zudem die Reihenfolge der darin enthaltenen Verweise (bei Relationen `<member type="way" ref=.../>` und bei Wegen `<nd ref=.../>`) wichtig, da diese die Abfolge der Elemente bestimmen.

4 OpenStreetMap Datenextrakte

Nachdem man die Struktur der OSM-Daten zur Beschreibung eines Postleitzahlengebietes kennt, ist es notwendig die XML-Dateien des Gebietes – hier also Deutschland – herunterzuladen. Dazu gibt es mehrere Möglichkeiten: Man kann selbst ausgewählte Daten exportieren, was allerdings nur bei kleinen Gebieten über die Export-Funktionalität der Webseite (http://www.openstreetmap.org/export) sinnvoll ist, da die Exporte schnell sehr groß werden und der Server sowohl eine zeitliche als auch größenmäßige Limitierung bei exportierbaren Daten vorgibt.
Bei großen Datenmengen ist es daher sinnvoller, auf vorhandene Datenexporte zurückzugreifen. Die sogenannte planet.osm Datei enthält beispielsweise alle Daten der Erde und wird täglich neu extrahiert, ist allerdings mehr als 730 GB groß. Aber auch kleinere Einheiten, z.B. für Deutschland können von verschiedenen Spiegelseiten heruntergeladen werden (http://wiki.openstreetmap.org/wiki/Planet.osm). Eine Exportdatei zum Gebiet von Deutschland ist dann allerdings immer noch mehr als 55 GB groß.
Will man sich auf einzelne Elemente, wie z.B. die Postleitzahlenregionen beschränken, so kann die Nutzung verschiedener Tools, wie z.B. der Overpass-API sinnvoll sein, da man hiermit gezielt Objekte auswählen kann. Nach Identifikation der geeigneten Map Features kann man dann eine Auswahl der Objekte treffen und diese über die Overpass-

API exportieren[4]. Das hier beschriebene Beispiel basiert auf einem Datenexport mit der Overpass-API als Datengrundlage, bei dem alle Postleitzahlengebiete-Relationen mit allen zugehörigen Wegen und Knoten extrahiert wurden[5], was den Datenumfang auf etwa 265 MB reduzierte.
Eine andere Möglichkeit besteht darin, basierend auf einem Download der Daten (vorzugsweise im Binärformat .pbf) von den erwähnten Spiegelseiten mithilfe des als Befehlszeilenprogramm für Windows, Linux und Mac zur Verfügung stehenden Tools OSMCONVERT[6], zuerst eine Konvertierung der Daten ins Binärformat .o5s vorzunehmen und die Daten anschließend mit dem ebenfalls frei verfügbaren Befehlszeilenprogramm OSMFILTER[7] auf die Postleitzahl-Relationen einzuschränken und daraus eine XML-Datei zu generieren.

5 SAS Map Data Sets

Um Informationen aus der OSM XML-Datei in ein SAS Map Data Set überführen zu können, muss auch die Struktur des SAS Map Data Set verstanden werden, die im Folgenden kurz erläutert werden soll.
In der Regel verwendet man in SAS zwei zusammengehörige Data Sets für die Erzeugung einer Karte: Einen Datensatz, das eigentliche Map Data Set, der die Koordinaten und eindeutige Zuordnung der Koordinaten zu einem Gebiet enthält und der dabei mindestens die Felder ID, X und Y sowie das optionale Feld SEGMENT umfasst (die Namen der Felder sind vorgegeben) sowie einen zweiten Datensatz, der zu diesen eindeutig durch die ID gekennzeichneten Gebieten weitere Informationen wie z.B. den geografischen Namen des Gebietes, aber auch die eigentlichen Auswertungsdaten zur Verfügung stellt (Response Data Set). Ziel ist es also, zwei Data Sets mit folgender Struktur zu erzeugen:

Tabelle 1: SAS Map Data Set (geografische Daten)

ID	SEGMENT	X	Y
1	1	5,768	52,125
1	1	5,763	52,132
..	..	..	..
1	1	5,768	52,125
2	1	6,123	51,723
2	1	6,125	51,720
…			

Im SAS Map Data Set liegt pro Gebiet (pro ID), eine Anzahl von Zeilen vor, die dieses Gebiet mit einer Abfolge von Punktkoordinaten (X und Y) beschreiben. Das optionale

[4] Besonders hilfreich ist hierbei die Webseite http://overpass-turbo.eu.
[5] Syntax der Abfrage: area[name='Deutschland'];rel[boundary=postal_code][note];(._;>;);out;
[6] http://wiki.openstreetmap.org/wiki/DE:Osmconvert
[7] http://wiki.openstreetmap.org/wiki/DE:Osmfilter

Feld SEGMENT wird dann benötigt, wenn ein Gebiet ein oder mehrere nicht zum Gebiet gehörige Polygone umschließt oder sich aus mehreren nicht zusammenhängenden Gebieten zusammensetzt.

Tabelle 2: Response Data Set (Bezeichnungen und Auswertungsdaten)

Id	Postleitzahl	Ortsbezeichnung	… ggf. weitere Variablen der eigentlichen Auswertung
1	34119	Kassel	
2	34121	Kassel	
…			

Ziel der weiteren Schritte ist es, diese beiden Tabellen zu erzeugen und mit Daten aus OSM zu füllen.

6 Einlesen von XML-Daten in SAS

In SAS besteht mit der XML Libname Engine eine Möglichkeit, XML-Dateien direkt in ein SAS Data Set einzulesen. Das mit dem XML-Format vorgegebene hierarchische Datenmodell muss dabei auf das zeilen- und spaltenbasierte Datenmodell einer SAS-Datentabelle übertragen werden. Dies geschieht über eine sogenannte MAP-Datei, in der ausgewählte Elemente der XML-Datei den Spalten bzw. Variablen der Datentabelle zugeordnet werden.
Wie bereits weiter oben beschrieben, bestehen die Postleitzahlengebiete, die in SAS eingelesen werden sollen, aus den drei Elementtypen Relation, Weg und Knoten. Aus allen drei Elementen müssen Informationen verwendet werden, will man daraus eine SAS Map und Response Data Set erzeugen. Da sich die Informationen dieser drei Elemente an ganz verschiedenen Stellen des XML-Dokumentes befinden, ist es notwendig die Elemente zunächst getrennt einzulesen. Im Folgenden soll die dazu erforderliche XML-Map näher erläutert werden.

6.1 Relationen

Um die Relationen in SAS einzulesen, benötigt man die `id` der Relation, da diese dem Postleitzahlengebiet einen eindeutigen Schlüssel gibt. Weiterhin werden die `tags` welche das Postleitzahlengebiet bezeichnen (im Beispiel: `<tag k="note" v="34119 Kassel"/>` sowie `<tag k="postal_code" v="34119"/>`), sowie alle Referenzen auf Wege benötigt, da diese dann auf die Informationen zum Umriss des Gebietes verweisen.

Die Struktur der Map-Datei legt dabei fest, welche XML-Tags eine SAS-Tabelle bilden sollen und wo sich im XML-Dokument die dazugehörigen Objekte befinden, die die Spalten dieser Tabelle bilden. Zur Referenzierung von einzelnen Elementen innerhalb des XML-Dokumentes bedient sich die SAS MAP-Syntax der sogenannten XPath-Syntax. Diese vom W3-Konsortium entwickelte Abfragesprache für XML-Dokumente

ist in SAS (mit einigen Einschränkungen) implementiert. Für Details dieser Sprache sei auf die Webseite des W3-Konsortiums[8] verwiesen.
Die nachfolgende XML-Map steuert die Zuordnung der in den Relationen vorhandenen Informationen zu Spalten bzw. Variablen in einem SAS Data Set.

```
<TABLE name="relation">
     <TABLE-PATH syntax="XPath">
          /osm/relation/member
     </TABLE-PATH>
     <COLUMN name="id" retain="YES">
          <PATH syntax="XPath">/osm/relation/@id</PATH>
          <TYPE>numeric</TYPE>
          <DATATYPE>integer</DATATYPE>
     </COLUMN>
     <COLUMN name="member_ref">
          <PATH syntax="XPath"> /osm/relation/member/@ref </PATH>
          <TYPE>numeric</TYPE>
          <DATATYPE>INTEGER</DATATYPE>
     </COLUMN>
     <COLUMN name="role">
          <PATH syntax="XPath"> /osm/relation/member/@role </PATH>
          <TYPE>character</TYPE>
          <DATATYPE>STRING</DATATYPE>
          <LENGTH>14</LENGTH>
     </COLUMN>
</TABLE>
```

In der Map werden dabei in der erste TABLE Anweisung über den `<TABLE-PATH>` (`/osm/relation/member`) die member-Objekte der relation-Ebene innerhalb des osm (osm = OpenStreetMap) Objektes gesucht, hier alle Verweise auf Wege innerhalb der Relation.
Attribute innerhalb eines XML-Tags lassen sich über das @-Symbol ansprechen. So wird in der ersten `COLUMN`-Anweisung (`/osm/relation/@id`) das Attribut `id` der Relation und in der nächsten `COLUMN`-Anweisung jeder Verweis (`ref`) auf einen Weg ausgelesen (`/osm/relation/member/@ref`).
Da die Anzahl der Wege innerhalb einer Relation unterschiedlich sein kann, muss man für jede Weg-Referenz eine Beobachtung vorsehen und diese mit der `id` der zugehörigen Relation versehen. Die `id` der Relation muss also so oft wiederholt werden, wie es Wege in der Relation gibt. Dieses Wiederholen kann man in der Syntax, ähnlich wie im SAS Data Step, durch eine `retain="YES"` erreichen.
Abschließend wird noch das Attribut `role` (`/osm/relation/member/@role`) erfasst. Dieses zeigt in OpenStreetMap ggf. an, ob sich innerhalb eines Gebietes (`role=outer`) ein Polygon befindet, das nicht zu diesem Gebiet gehört (`role=inner`) – und sinngemäß ein Loch im Gebiet bildet. Für das oben dargestellte Beispiel der Postleitzahl

[8] https://www.w3.org/TR/xpath20

34119 der Stadt Kassel wird für die Relation mit der `id=1189969` über die XML-Map eine Tabelle wie nachfolgend dargestellt erzeugt.

Tabelle 3: Relation mit Verweisen auf die zugehörigen Wege und Rolle

id	ref	role
1189969	79123010	outer
1189969	453136782	outer
1189969	79123005	outer
1189969	79123008	outer
1189969	79123023	outer
1189969	79122997	outer
1189969	79123006	outer

Ein zweiter MAP-Abschnitt ist notwendig, da man die in der Relation ebenfalls vorhandenen Informationen zum Postleitzahlengebiet in ein separates Data Set einlesen will.

```
<TABLE name="data">
     <TABLE-PATH syntax="XPath">
      /osm/relation
  </TABLE-PATH>
  <COLUMN name="id" retain="YES">
     <PATH syntax="XPath"> /osm/relation/@id </PATH>
     <TYPE>numeric</TYPE>
     <DATATYPE>integer</DATATYPE>
     </COLUMN>
     <COLUMN name="postal_code">
     <PATH syntax="XPath">
          /osm/relation/tag/@v[@k="postal_code"]</PATH>
     <TYPE>character</TYPE>
     <DATATYPE>STRING</DATATYPE>
     <LENGTH>6</LENGTH>
  </COLUMN>
  <COLUMN name="note">
     <PATH syntax="XPath"> /osm/relation/tag/@v[@k="note"]
               </PATH>
     <TYPE>character</TYPE>
     <DATATYPE>STRING</DATATYPE>
     <LENGTH>1000</LENGTH>
     </COLUMN>
</TABLE>
```

Die entstehende Tabelle hat für 34119 Kassel genau eine Zeile:

Tabelle 4: Postleitzahl und Ortsangabe zur Relation

id	postal_code	note
1189969	34119	34119 Kassel

6.2 Wege

Analog zu den Relationen wird auch für die Wege ein XML-Map Abschnitt benötigt.

```
<TABLE name="ways">
     <TABLE-PATH syntax="XPath">
     /osm/way/nd
  </TABLE-PATH>
  <COLUMN name="way_id" retain="YES">
     <PATH syntax="XPath"> /osm/way/@id </PATH>
     <TYPE>numeric</TYPE>
     <DATATYPE>integer</DATATYPE>
     </COLUMN>
     <COLUMN name="nd_ref">
     <PATH syntax="XPath"> /osm/way/nd/@ref </PATH>
     <TYPE>numeric</TYPE>
     <DATATYPE>integer</DATATYPE>
     </COLUMN>
</TABLE>
```

Für den ersten Weg der Relation aus dem Beispiel sieht die Tabelle, die erzeugt wird wie folgt aus (gekürzt):

Tabelle 5: Verweise auf Knoten eines Weges

way_id	nd_ref
79123010	924504450
79123010	924504256
79123010	924504137
79123010	925551353
…	

6.3 Nodes

Der XML-Map Abschnitt für das Einlesen der Knoten sieht folgendermaßen aus:

```
     <TABLE name="nodes">
        <TABLE-PATH syntax="XPath">
           /osm/node
         </TABLE-PATH>
        <COLUMN name="node_id" retain="YES">
          <PATH syntax="XPath"> /osm/node/@id </PATH>
            <TYPE>numeric</TYPE>
         </COLUMN>
          <COLUMN name="lon">
          <PATH syntax="XPath"> /osm/node/@lon </PATH>
            <TYPE>numeric</TYPE>
            <DATATYPE>double</DATATYPE>
         </COLUMN>
           <COLUMN name="lat">
```

```
          <PATH syntax="XPath"> /osm/node/@lat </PATH>
            <TYPE>numeric</TYPE>
            <DATATYPE>double</DATATYPE>
          </COLUMN>
      </TABLE>
```

Über die oben dargestellte Map-Syntax kann die nachfolgende Tabelle erzeugt werden (gekürzt).

Tabelle 6: Koordinaten der Knoten eines Weges

node_id	lon	lat
924504450	9,46329	51,325323
924504256	9,461618	51,323278
924504137	9,460376	51,321673
925551353	9,458988	51,321325
…	..	…

Das eigentliche Einlesen der XML-Dateien und die Erzeugung der Data Sets erfolgt nach Erstellung einer XML-Map, über den SAS Data Step. Die XML-Map wird dabei über `libname` und `filename` referenziert. Der Teil der XML-Map der auf die Tabelleneigenschaften der Tabelle verweist die erzeugt werden soll, wird über die Punkt-Notation, also `.Tabellenname` angesprochen (im Beispiel: die Tabelle `relation`).

```
filename OXMLIn "Pfad_zur_OSM_XML_Datei" ;
filename OXMLMap "Pfad_zur_XML_Map_Datei";
libname OXMLIn xml xmlmap=OXMLMap;

data relation;
     set OXMLIn.relation; *Verweis auf die Tabelle in der XML-Map;
     retain rel_ref_no; if id ne lag(id) then rel_ref_no=0;
     rel_ref_no=rel_ref_no+1; *Nummerierung ;
     format rel_ref_no 14.;
run;
```

Im Beispiel wird für jede Referenz auf einen Weg pro Relation eine fortlaufende Nummerierung erzeugt, um die Reihenfolge der Elemente ab dem ersten Einlesevorgang nachvollziehbar zu halten, da diese die Darstellung auf der Karte wesentlich mitbestimmt.
Analog dazu kann man auch die Tabellen für die Wege sowie die Knoten erzeugen.
Sind dann die Informationen zu Wegen und Knoten eingelesen, müssen aus den einzelnen Data Sets die für die Erstellung des SAS Map Data Set notwendigen Informationen zusammengeführt werden. Dazu werden die Tabellen 3, 5 und 6 über einen PROC SQL-Befehl zu einer Tabelle verknüpft, welche dann als Grundlage für den SAS Map Data Set dient. Daten aus Tabelle 4 werden als Response Data Set verwendet, in dem pro Postleitzahlengebiet eine Zeile für den zugehörigen Primärschlüssel (ID) vorliegt.

6.4 Fallstricke bei der Erzeugung des SAS Map Data Sets

Die GMAP-Prozedur in SAS/GRAPH erwartet, dass die einzelnen Koordinaten eines geschlossenen Gebietes in korrekter Reihenfolge im Map Data Set vorliegen. Da sich in OSM die Grenzen eines Postleitzahlengebiets aus Abfolgen von unterschiedlichen Wegen zusammensetzen, kann es vorkommen, dass die Reihenfolge der Punkte von Weg zu Weg wechseln kann und dann zum Teil nicht der Reihenfolge entspricht, wie sie durch die GMAP-Prozedur erwartet wird. Es kann also z.B. sein, dass ein Weg von Ost nach West gezeichnet ist, der nächste Weg aber von West nach Ost. In diesem Fall müssen die Knoten des einen Weges umsortiert werden.
Die Notwendigkeit einer neuen Sortierung muss dabei schrittweise entschieden werden, basierend darauf, ob der letzte Knoten eines Weges gleich dem ersten Knoten des Folgeweges ist. Auch im oben verwendeten Beispiel 34119 Kassel war der letzte Weg zunächst in seiner Knoten-Reihenfolge umgekehrt, was zu einer falschen Darstellung wie in Abbildung 4 führt. Man sieht, dass die Darstellung am oberen Rand eine zusätzliche Linie aufweist, die dadurch entsteht, dass die GMAP-Prozedur den letzten Knoten des letzten Weges zuerst ansteuert und sich dann über die restlichen Knoten zum Ausgangspunkt zurück bewegt.

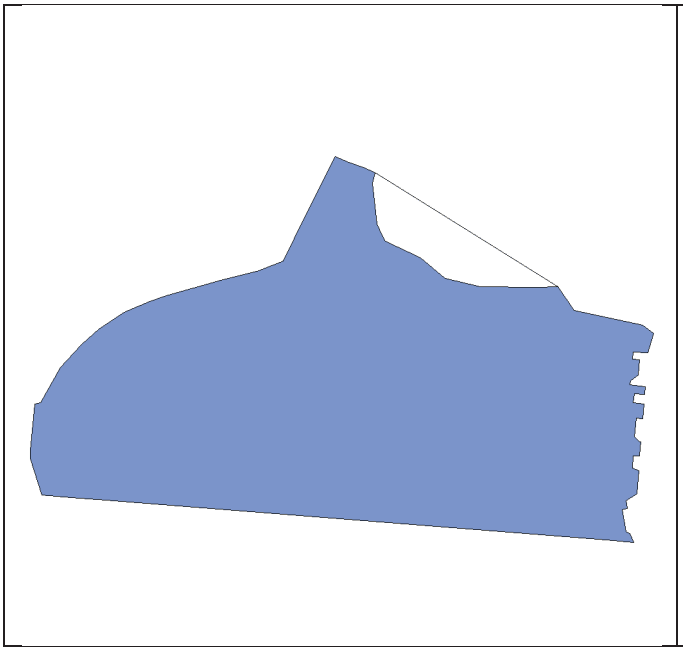

Abbildung 4: Postleitzahlengebiet 34119 Kassel mit falscher Reihenfolge der Knoten des letzten Weges

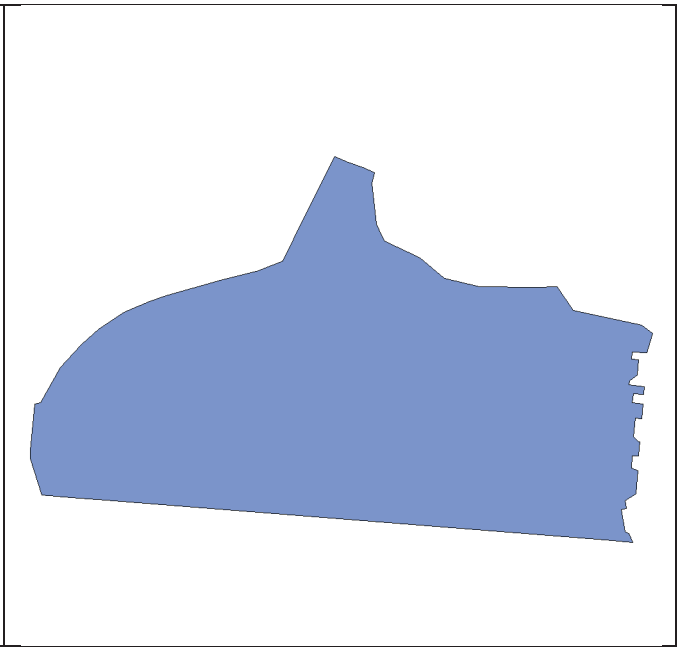

Abbildung 5: Postleitzahlengebiet 34119 Kassel in korrekter Darstellung

Weiterhin gibt es auch den Fall, dass ein Postleitzahlengebiet die Fläche nicht komplett ausfüllt, sondern innerhalb des Gebietes die Enklave eines anderen Postleitzahlengebietes existiert. In OSM sind diese Enklaven durch einen oder mehrere Wege gekennzeichnet, die als Attribut `role = inner` aufweisen (vgl.**Tabelle 3**Tabelle 3). Als Beispiel wird in Abbildung 6 und Abbildung 7 das Postleitzahlgebiet 53881 Euskirchen dargestellt, das als Enklave ein anderes Postleitzahlengebiet im Inneren aufweist.

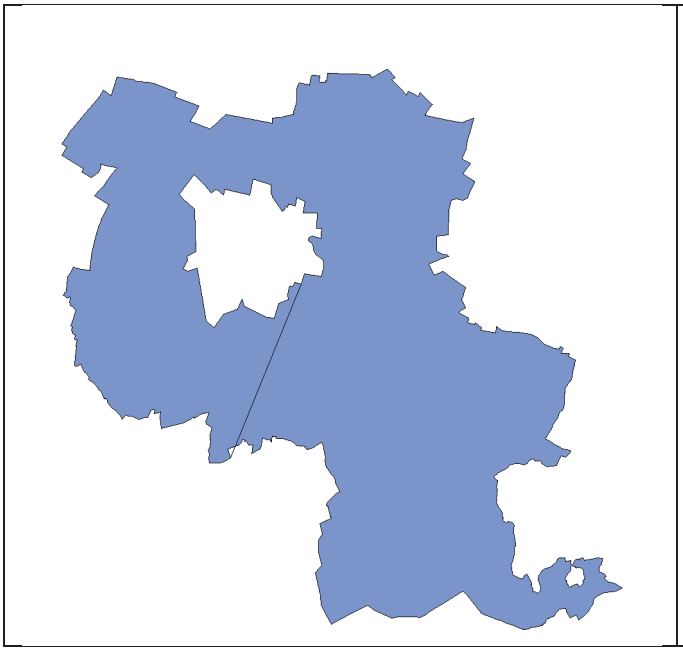

Abbildung 6: Postleitzahlengebiet 53881 Euskirchen mit Enklave in falscher Darstellung

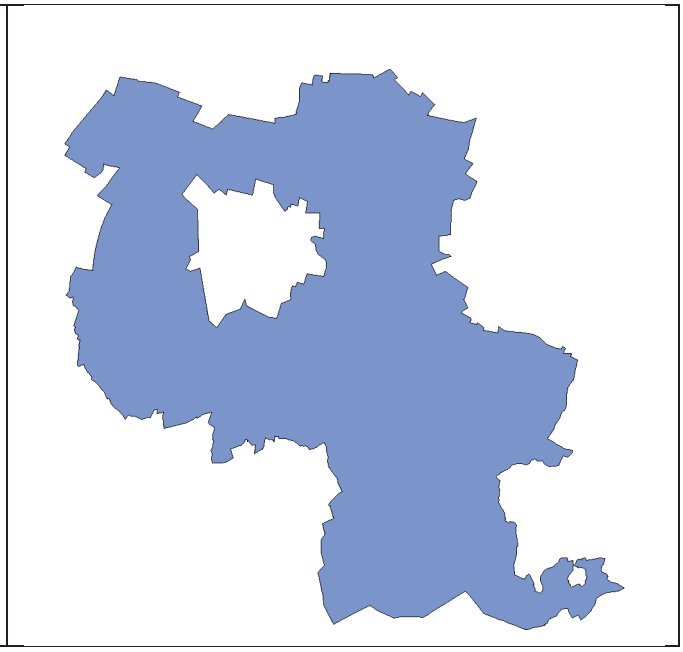

Abbildung 7: Postleitzahlengebiet 53881 Euskirchen mit Enklave in korrekter Darstellung

Man sieht in Abbildung 6, dass die GMAP-Prozedur zunächst eine Verbindungslinie zwischen dem äußeren und inneren Gebiet erzeugt. Um dies zu verhindern, muss im SAS Map Data Set nach dem letzten Knoten des äußeren Polygons eine Zeile mit gleichbleibender ID und einem MISSING in den Spalten SEGMENT, X und Y erzeugt werden (vgl.Tabelle 7).

Tabelle 7: Repräsentation von Enklaven in SAS Map Data Sets

ID	SEGMENT	X	Y
158144	1	6,77724	50,60650
158144	1	6,77713	50,60649
158144	1	6,77746	50,60567
158144	1	6,77820	50,60410
158144	.	.	.
158144	1	6,80646	50,64829
158144	1	6,80782	50,65103
158144	1	6,81434	50,65025

Neben diesem Fall gibt es auch noch Gebietseinheiten, die sich aus mehreren einzelnen Flächen zusammensetzen. So bildet zum Beispiel 56335 Nastätten ein Postleitzahlengebiet, das aus mehreren separaten Gebieten zusammengesetzt ist.

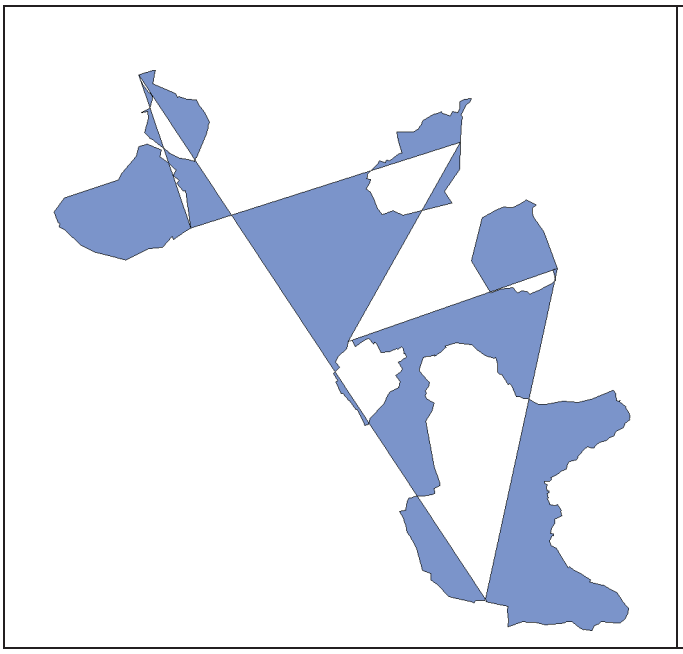

Abbildung 8: 56335 Nastätten in falscher Darstellung

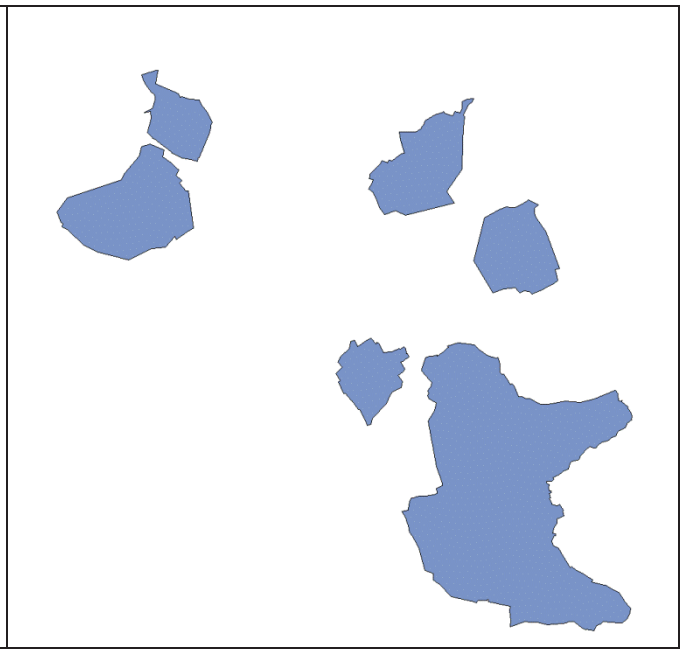

Abbildung 9: 56335 Nastätten in korrekter Darstellung

In diesem Fall erwartet die GMAP-Prozedur für jede Fläche eine eigene Segmentkennung innerhalb der unveränderten ID, wie in Tabelle 8 dargestellt.

Tabelle 8: Repräsentation von zusammengesetzten Gebieten in SAS Map Data Sets

ID	SEGMENT	X	Y
1130664	1	7,86254	50,15507
1130664	1	7,86221	50,15516
1130664	2	7,88739	50,22876
1130664	2	7,88570	50,22844
1130664	2	7,88646	50,22699
..	..	..	..
1130664	3	7,81479	50,21250
..	..	..	..

Dieser Fall ist schwieriger zu identifizieren, da keine gesonderte Kennung in der XML-Datei vorhanden ist. Einziger Anhaltspunkt ist ein Knoten mit identischer Knoten-ID am Ende wie zu Beginn eines Weges. Man muss also die Knoten-ID des ersten Knotens eines Polygons mithilfe eines `retain` Befehls fortschreiben, um diese Knoten-ID mit der Knoten-ID am Ende eines Weges vergleichen zu können. Sind die Knoten-IDs identisch, so endet das erste Polygon des Postleitzahlgebietes und man kennzeichnet dies, indem man beim nächsten Polygon dieser Postleitzahl die SEGMENT-Nummer um eins erhöht.

7 Erstellung der Grafik

Nachdem mit den zuvor beschriebenen Schritten die Daten zu Deutschland eingelesen und durch weitere Schritte die Wege der Polygone der einzelnen Gebiete in die richtige Reihenfolge gebracht wurden sowie die Enklaven identifiziert und mehrteilige Postleitzahlengebiete korrekt durchnummeriert wurden, ist noch ein letzte Schritt zu vollziehen: Die Umwandlung der in OSM gemäß WGS 84 vorliegenden geografischen Koordinaten in projizierte Koordinaten, die von der GMAP-Prozedur in SAS korrekt dargestellt werden können. Dazu verwendet man die GPROJECT-Prozedur von SAS/GRAPH in folgender Weise:

```
proc gproject data=plzmap DEG EASTLONG DUPOK
   out=postleitzahlmap;
   id id;
run;
```

Wichtig ist es in diesem Zusammenhang, mit den Optionen DEG EASTLONG und DUPOK zu arbeiten, welche der Prozedur angeben, dass es sich im Ausgangsdatensatz um Angaben in Längen- und Breitengraden handelt (DEG), diese entgegen der erwarteten Richtung negative Gradzahlen aufweisen (EASTLONG; Standard: Westlich des Nullmeridians werden positive Werte angegeben, im Falle von WGS 84 sind diese Negativ) sowie aufeinanderfolgende Knoten mit gleichen Koordinaten nicht zu löschen (DUPOK). Die letzte Option ist nicht unbedingt notwendig, nachdem man die oben genannten Probleme gelöst hat, aber leistete während des Debuggings gute Dienste.

Mit den entstandenen Datensätzen kann dann, mit dem Aufruf der GMAP-Prozedur eine Kartendarstellung (siehe Abbildung 10) erzeugt werden.

```
PROC GMAP
     map=postleitzahlmap * der Map Data Set;
     data=data all;      * der Response Data Set;
     id id;              * die Verknüpfungsvariable;
     choro plz1 /nolegend *plz1: enthält die ersten Ziffer der PLZ;
RUN;QUIT;
```

Gleiche Farben kennzeichnen Gebiete mit derselben ersten Ziffer der Postleitzahl, was über die zusätzlich angelegte Variable plz1 gesteuert wird, die diese Ziffer enthält. Die im Süden Bayerns dargestellten „Löcher“ in der Karte sind keine Einlese-, oder Darstellungsfehler, sondern Gebiete, die in OSM nicht Teil eines Postleitzahlengebietes sind – es handelt sich um den Ammersee und den Starnberger See. Insgesamt enthält das SAS Map Data Set etwa 5,3 Millionen Zeilen mit den einzelnen Koordinaten der insgesamt 8.183 Postleitzahlengebiete (Stand: Ende 2016) in Deutschland.

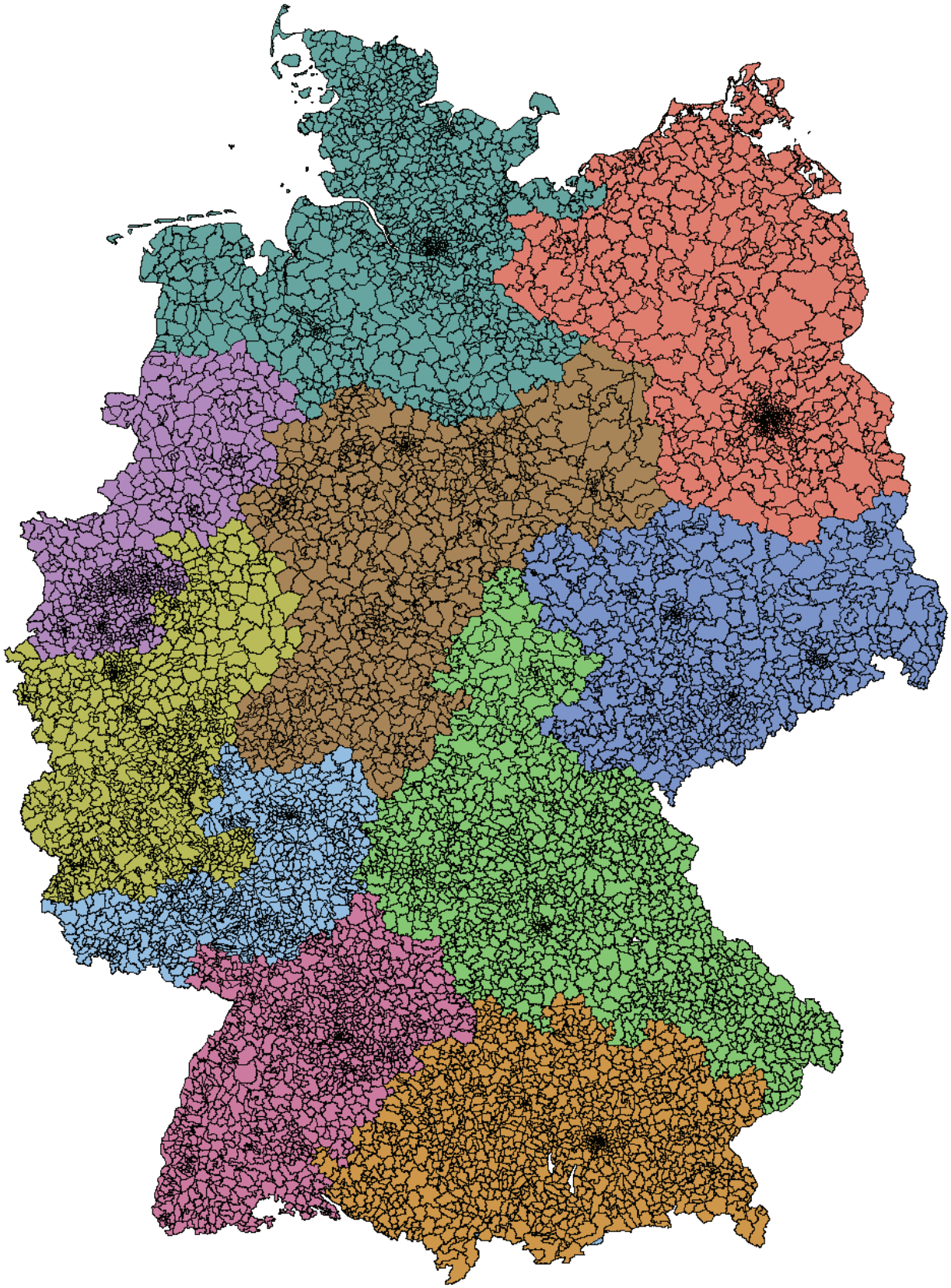

Abbildung 10: Darstellung der Postleitzahlengebiete von Deutschland

8 Fazit

Die Verwendung von OSM-Daten als Grundlage von Kartendarstellungen in SAS ist ohne weiteres möglich. Der wesentliche Aufwand ist dabei nicht die Übertragung der Daten in SAS, sondern die damit verbundene Durchdringung der Datenstrukturen von OSM – Stichwort: Map Features – sowie die Umstrukturierung der Daten in den beschriebenen Sonderfällen.

Weiterhin sollte man beachten, dass die Vollständigkeit der Daten in OSM davon abhängt, ob es genügend Beitragende mit dem gemeinsamen Interesse gibt, bestimmte geografische Eigenschaften zu kartieren. Häufig werden für bestimmte Themen Projekte ins Leben gerufen, die sich einem Thema widmen und hier auch die Kartierung sowie deren und Qualitätssicherung Vereinheitlichung vorantreiben[9].

[9] Weitere Informationen dazu finden sich auf: http://wiki.openstreetmap.org/