Nachtrag zum `tectr`-Exposé

Samuel Lippl

25 Januar 2018

Operatoren und Compounds - ein Beispiel

Zur Veranschaulichung der Definition

$$Op : \mathcal{C} \rightarrow V, V(C^0) = NA$$

werde ich Daten der Weltbank und der UCDP heranziehen. Die Datenbank der UCDP gibt es in verschiedenen Aggregationen, z. B. auf Koordinaten-Tagesebene. (`best` bezieht sich auf den Best Estimate der Toten.)

```
load("data/GED 17.1.Rdata")
library(sf)
```

```
## Linking to GEOS 3.6.1, GDAL 2.2.0, proj.4 4.9.3
```

```
ged171 <- ged171 %>%
  st_as_sf(coords = c("coords.x1", "coords.x2"))
```

```
## Loading required package: sp
```

```
head(ged171 %>%
       select(year, dyad_name,
              date_start, date_end, best))
```

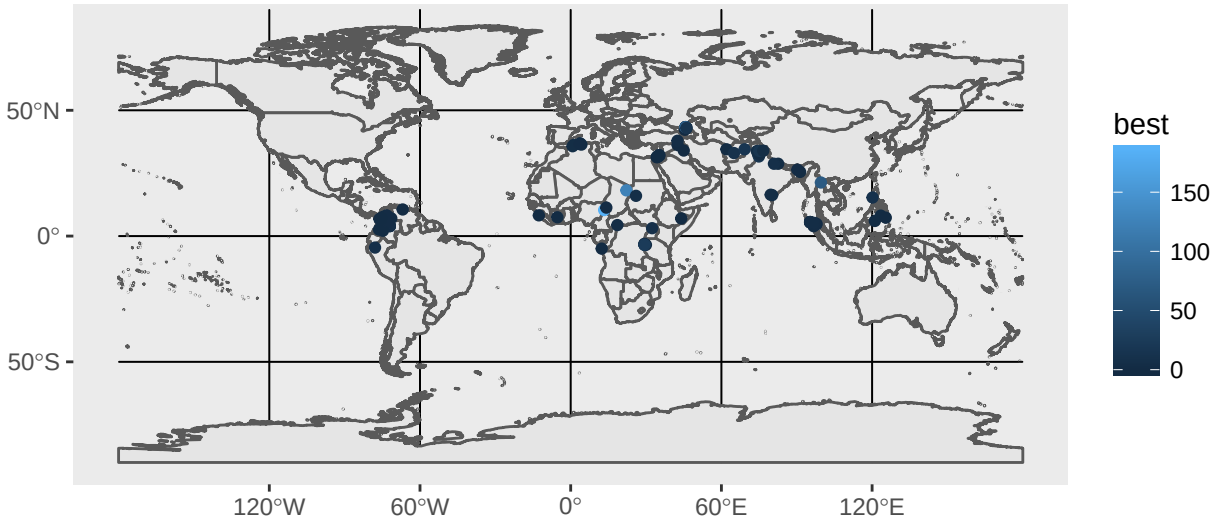
```
## Simple feature collection with 6 features and 5 fields
## geometry type:  POINT
## dimension:      XY
## bbox:           xmin: 42.58333 ymin: 32.44179 xmax: 75.10929 ymax: 36.16667
## epsg (SRID):    4326
## proj4string:    +proj=longlat +datum=WGS84 +no_defs
##   year                                  dyad_name date_start   date_end
## 1 2013                      Government of Iraq - IS 2013-08-06 2013-08-06
## 2 2014 Government of India - Government of Pakistan 2014-01-31 2014-01-31
## 3 2014 Government of India - Government of Pakistan 2014-01-31 2014-01-31
## 4 2014                      Government of Iraq - IS 2014-02-22 2014-02-22
## 5 2014          Government of Afghanistan - Taleban 2014-05-06 2014-05-06
## 6 2014          Government of Afghanistan - Taleban 2014-07-11 2014-07-12
##   best                    geometry
## 1    0               POINT (45 34)
## 2    1  POINT (75.057907 32.44179)
## 3    1   POINT (75.10929 32.57523)
## 4    0 POINT (42.583333 36.166667)
## 5    1             POINT (62 34.5)
## 6   11               POINT (65 33)
```

Die Koordinaten der Spalte `geometry` sind ein Beispiel für einen Compound im Kontext von Fragilität; auch die Zeitpunkte (bzw. Zeitintervalle) sind Compounds. Compounds können also multidimensional sein; in diesem Fall sind sie ein geordnetes Paar `(spatial, time)`. Ich werde mich im Folgenden zur Veranschaulichung auf die Koordinatendimension beschränken. Haben wir nur solche Daten, so ist die Visualisierung recht einfach: wir nehmen uns zur Referenz noch eine geeignete Datenbank, in der Ländergrenzen aufgeführt sind. Alles andere erledigt `ggplot`:

```
cou <- st_read("data/Countries Shapefile/ne_10m_admin_0_countries.shx")
```

```
## Reading layer `ne_10m_admin_0_countries' from data source `C:\Users\samue\Documents\Studium\Statistik
## Simple feature collection with 255 features and 71 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: -180 ymin: -90 xmax: 180 ymax: 83.6341
## epsg (SRID):    4326
## proj4string:    +proj=longlat +datum=WGS84 +no_defs
```

```
ggplot(cou) +
  geom_sf() +
  geom_sf(data = head(ged171, 100),
          mapping = aes(colour = best))
```



Ein Beispiel für einen Operator wäre nun der Best Estimate der Toten `deaths`. In dem Fall wäre die Bildmenge des Operator $V = \mathbb{N}$, weil die absolute Zahl der Toten ja immer eine ganze Zahl ist. Wir könnten auch einfach $V = \mathbb{R}$ setzen, da ja $\mathbb{N} \subset \mathbb{R}$ und nicht die ganze Bildmenge vom Operator abgedeckt werden muss.
V ist also die Menge von Werten, mit denen unser Operator eine Aussage über einen Compound trifft. Meistens ist $V \subseteq \mathbb{R}$, allerdings könnte ein Operator auch eine Funktion oder etwas anderes Exotischeres liefern.
Wir können nun also abfragen, was `deaths` für eine bestimmte Koordinate und einen bestimmten Zeitrahmen ausgibt. Entspricht dies einer Zeile in der Quelle, gibt `deaths` diesen Wert zurück, ansonsten `NA`.
Das ist ein wenig unpraktisch. Eine Region wie Afghanistan wäre ebenfalls ein Compound und wir haben ja durchaus Information darüber gegeben, wie viele Menschen in diesem Land gestorben sind. Wir könnten also eine Erweiterung von `deaths` definieren, die wir `death_ext` nennen und uns für jede Region die Zahl der darin gestorbenen Menschen angibt. `death_ext` ist nun also ein Internal Operator, der in seinen Werten

von `deaths` abhängt. Damit liegt eine klare Trennung zwischen den eingelesenen Daten und den daraus gezogenen Schlüssen vor.
Die UCDP hat sehr nutzerfreundliche räumliche Daten. Die meisten Datenbanken sind unzugänglicher. Beispielsweise gibt es in der Weltbank einzelne Ländernamen bzw. dreistellige Codes und Werte für einzelne Länder und Jahre – keine Koordinaten oder Koordinatenpolygone. Dazu sind die Daten nur als csv gegeben, das ein bisschen Vorverarbeitung braucht. Als einfaches Beispiel habe ich hier die Bevölkerung genommen. Die Funktion `wb_read` kümmert sich um das Einlesen und könnte in `tectr` mit dem entsprechenden Operator assoziiert werden, um diesen Vorgang zu vereinfachen.

```
wb_read <- function(path, valname) {
  possible_files <- dir(path = path)
  which_files <- possible_files %>%
    grepl("Metadata", ., fixed = TRUE)
  file <- possible_files[!which_files]
  if(length(file) != 1) stop("Weird directory!")
  read.csv(paste(path, file, sep = "/"), skip = 4, header = TRUE) %>%
    select(-X) %>%
    gather(year, Value, starts_with("X"), na.rm = TRUE) %>%
    mutate(year = parse_number(year),
           code = Country.Code,
           name = valname) %>%
    spread(name, Value) %>%
    select(year, code, !!valname)
}
pop <- wb_read("data/Worldbank/Population", "pop")
head(pop)
```

```
##   year code     pop
## 1 1960  AFG 8996351
## 2 1961  AFG 9166764
## 3 1962  AFG 9345868
## 4 1963  AFG 9533954
## 5 1964  AFG 9731361
## 6 1965  AFG 9938414
```

Compounds können also auch Länder wie AFG, also Afghanistan, und ganze Jahre sein. In diesem Fall ist unsere Bevölkerung ein Operator, der, falls wir ein Land eingeben und er entsprechende Daten hat, diese Daten, die wieder in $V = \mathbb{R}$ liegen, zurückgibt. `AFG` ist hierbei eine Bezeichnung für das zugrundeliegende Koordinatenpolygon. Geben wir ein Land ein, für das die Weltbank keine Daten bietet, gibt der Operator `NA` zurück. Auch wenn wir eine Koordinate eingeben, gibt der Operator `NA` zurück. Auch von diesem Operator können wir eine Erweiterung definieren, die beispielsweise für die zusammen genommenen Länder Afghanistan und Deutschland, dargestellt durch die Syntax `AFG + GER` ihre Bevölkerung `pop(AFG + GER) = pop(AFG) + pop(GER)` zurückgibt. Wohlgemerkt müssen wir für `AFG + GER` kein Koordinatenpolygon speichern, da dieses aus den Koordinatenpolygonen `AFG` und `GER` herleitbar ist.
Die Operatoren und Compounds bieten in ihrer Kombination also insbesondere ein flexibles Framework, das mit Daten unterschiedlicher Auflösung umgehen kann und diese auf ein gemeinsames Level bringt, ohne detailliertere Informationen zu vernachlässigen (das ist zumindest das Ziel). Speziell würde sich hier im Kontext räumlicher Information eine Funktion anbieten, die bei der Assoziation der entsprechenden Regionen mit den Ländernamen Hilfe leistet. Weder die Bezeichnungen noch die Länder selbst sind nämlich einheitlich - beispielsweise ist Französisch-Guyana manchmal eine eigene Observation, während es in der Weltbank zu Frankreich zu zählen ist.

Zielsetzung

Weil ich schlecht abschätzen kann, wie gut ich mit dem Programmieren vorankommen werde, würde ich meine Zielsetzung in zwei Abstufungen darstellen: “Deliverables”, die am Ende verbindlich Teil der abgegebenen Arbeit sein werden (im Rahmen möglicher Änderungen im Sinne bessere Implementierungen) und weitere mögliche Zielen (die sich natürlich auch während der Arbeit ergeben könnten). Die tatsächliche Arbeit zur Fragilität sehe ich nicht als Teil dieser Bereiche. Meines Erachtens würde diese ein gutes Beispiel bereitstellen, um das Paket zu motivieren und zu erklären, aber die Entwicklung des Fragilitätsindikators ist Gegenstand meiner Hilfskraftstelle und daher meiner Meinung nach nur im Sinne einer Veranschaulichung Teil der Bachelorarbeit.

Deliverables

- die abstrakten Klassen der Contents, Operators und Compounds mit der Möglichkeit diese zu personalisieren, durch Methoden wie `read` oder `visualize`. Hierunter fallen darüber hinaus z. B.
  - ein sinnvoller Umgang mit sich aktualisierenden Daten
  - die Snapshots, mit denen sich die Konstruktion von Operatoren durch Regression nachverfolgen und flexibel verändern lässt
- Default-Klassen und -Methoden, mit denen sich einfache Strukturen unkompliziert umsetzen lassen. Auch das Minimalergebnis sollte für die Anwendung nützlich sein.
- die Erstellung entsprechender Visualisierungsmethoden, vermutlich mithilfe einer Shiny-App
- ausführliche Dokumentation des Pakets

Weitere Gebiete

- Umgang mit räumlichen Variablen, insbesondere solchen, die Länder oder Provinzen repräsentieren
- bei Bedarf Schreiben einzelner Funktionen in C++ mithilfe von Rcpp (würde mir willkommene Gelegenheit geben, dieses Paket kennenzulernen und C++ zu üben)
- Automatischer Download bei aktualisierten Internetquellen wäre ein praktisches Feature