

Using KOR.addrlink

Daniel Schürmann

February 29, 2024

1 Introduction

Consider a data set with semi-structured address data, e.g. street and house number as a concatenated string, wrongly spelled street names or non-existing house numbers. This data set (referred to as `df_match`) should be mapped to a complete list of valid addresses within the given municipality. The latter data set is called `df_ref` and may include further information like coordinates of addresses or district information. `KOR.addrlink` tries to solve this problem specifically for German municipalities as the package is specialized on German address schemes.

2 Reference data

First, a complete list of reference addresses (`df_ref`) is needed. An example data.frame named "Adressen" is shown below.

```
> library(KOR.addrlink)
> Adressen[c(sample(which(is.na(Adressen$HNRZ)), 4),
+             sample(which(!is.na(Adressen$HNRZ)), 2)),]
      STRNAME STRSL HNR HNRZ      RW      HW UBZ
69266  EMIL-FIGGE-STRAÙE 70951  21 <NA> 391077.7 5705865 642
32637      LANGE STRAÙE 76269  13 <NA> 392617.9 5708017 021
45713  SCHÜRUFERSTRAÙE 72770 277 <NA> 399291.7 5705557 413
86680      LIBORISTRAÙE 71998  36 <NA> 396753.8 5708098 073
27524 IMMANUEL-KANT-STRAÙE 71616  7    C 385367.6 5706739 734
87693      LÜNINCKWEG 76294  5    B 400334.3 5706476 411
```

The columns used for the matching procedure are `STRNAME` (street name), `HNR` (house number) and `HNRZ` (additional letter). This vignette illustrates the merging workflow on two sample data sets called `df1` and `df2`.

3 Example 1

df1 has address information in columns gross_strasse and housnr. The columns Var1 and Var2 provide non-address related information about the individuals. Row 1183 shows that the column hausnr needs to be split into house number and additional letter before addresses can be matched. The function split_number is provided for that task.

```
> df1[1180:(1183+6),]

      gross_strasse housnr Var1 Var2
1180      AZALEENWEG    39   4   B
1181      BRECHTSTRAE    12   2   B
1182          BURGWALL     6   1   A
1183      AM KRAFTWERK    5c   4   B
1184          IM DEFDAHL  355   1   A
1185 HILGENSTOCKSTRAE     4   1   A
1186      HERIBERTSTRAE     1   4   A
1187          WYNEKENWEG     1   3   B
1188          OEVERGUNNE    89   4   B
1189          GRUNEWALD    63   3   B
```

split_number takes hausnr and creates a data.frame with columns "Hausnummer" (house number) and "Hausnummernzusatz" (additional letter).

```
> df1 <- cbind(df1, split_number(df1$hausnr))
> df1[1180:(1183+6),]

      gross_strasse housnr Var1 Var2 Hausnummer Hausnummernzusatz
1180      AZALEENWEG    39   4   B          39          <NA>
1181      BRECHTSTRAE    12   2   B          12          <NA>
1182          BURGWALL     6   1   A           6          <NA>
1183      AM KRAFTWERK    5c   4   B           5             c
1184          IM DEFDAHL  355   1   A         355          <NA>
1185 HILGENSTOCKSTRAE     4   1   A           4          <NA>
1186      HERIBERTSTRAE     1   4   A           1          <NA>
1187          WYNEKENWEG     1   3   B           1          <NA>
1188          OEVERGUNNE    89   4   B          89          <NA>
1189          GRUNEWALD    63   3   B          63          <NA>
```

addrlink merges the two data sets. For both data sets, the columns referring to street name, house number and additional letter need to be specified in exactly that order (parameter col_ref and col_match).

```
> # column hausnr is no longer needed
> df1 <- within(df1, rm(hausnr))
> df1_matched <- addrlink(df_ref = Adressen,
```

```
+      col_ref = c("STRNAME", "HNR", "HNRZ"),
+      df_match = df1,
+      col_match = c("gross_strasse", "Hausnummer", "Hausnummernzusatz"))
```

The result is a list with two data.frames

- ret: The merged data set
- QA: Indicators showing the match quality

```
> head(df1_matched$ret)
```

	gross_strasse	Var1	Var2	Hausnummer	Hausnummernzusatz	STRNAME	STRSL
5876	AALBECKESTRAÙE	2	C	9	<NA>	AALBECKESTRAÙE	76567
3115	ABBOWEG	3	B	11	<NA>	ABBOWEG	70002
8760	ABBOWEG	3	C	20	<NA>	ABBOWEG	70002
2962	ABBOWEG	4	B	7	<NA>	ABBOWEG	70002
5554	ABTEISTRAÙE	1	C	10	<NA>	ABTEISTRAÙE	70003
3110	ABTEISTRAÙE	2	A	28	<NA>	ABTEISTRAÙE	70003

	HNR	HNRZ	RW	HW	UBZ
5876	9	<NA>	399330.4	5709633	324
3115	11	<NA>	387016.0	5708290	843
8760	20	<NA>	387134.8	5708466	843
2962	7	<NA>	386999.3	5708259	843
5554	10	<NA>	400342.6	5705684	414
3110	28	<NA>	400423.6	5705602	414

```
> table(df1_matched$QA$qAddress)
```

1	2	3	4
9670	72	157	101

qAddress states the stage within the matching procedure that yielded the match. Out of the 10000 records, 9670 could be merged directly. 72 had a valid street name, but an invalid house number. 157 records had (possibly) misspelled street names and 101 records could not be matched at all.

4 Example 2

The second data set has a single column "Adresse", which includes street names and house numbers. Thus, this column needs to be split by the function `split_address`.

```
> head(within(df2, Adresse <- trimws(Adresse)))
```

	Adresse	Var1	Var2
1	Wittbräucker Str. 584	4	B
2	Dünnebecke 72	4	A

```

3      Hermannstr. 4-6    2    C
4      Wenkerstr. 10     4    C
5      Baaderweg 11     3    C
6      Erfurter Str. 22-24 4    C

```

split_number creates a data.frame with columns "Strasse" (street) "Hausnummer" (house number) and "Hausnummernzusatz" (additional letter) from the column "Adresse".

```

> df2 <- cbind(df2, split_address(df2$Adresse))
> within(df2, Adresse <- trimws(Adresse))[23:(23+6),]

```

	Adresse	Var1	Var2	Strasse	Hausnummer
23	Albert-Schweitzer-Weg 14	1	A	Albert-Schweitzer-Weg	14
24	Am Bönner 36	2	C	Am Bönner	36
25	Dorstfelder Hellweg 90	3	C	Dorstfelder Hellweg	90
26	Vieselerhofstr. 26e	2	B	Vieselerhofstr.	26
27	Rosental 18	2	A	Rosental	18
28	Schwerter Str. 385	2	A	Schwerter Str.	385
29	Bockenfelder Str. 243	1	A	Bockenfelder Str.	243
	Hausnummernzusatz				
23	<NA>				
24	<NA>				
25	<NA>				
26	E				
27	<NA>				
28	<NA>				
29	<NA>				

Again, addrlink merges the two data sets. The parameter fuzzy_threshold sets the threshold for fuzzy matching of misspelled street names. A value of 1 means no fuzzy matching and 0 means forced fuzzy matches for all records. If a street name could be matched, but the provided house number does not exist, addrlink may randomly assign a valid house number to that record. A seed is always set to ensure reproducibility. Customization is possible via the parameter seed.

```

> # column Adresse is no longer needed
> df2 <- within(df2, rm(Adresse))
> df2_matched <- addrlink(df_ref = Adressen,
+   col_ref = c("STRNAME", "HNR", "HNRZ"),
+   df_match = df2,
+   col_match = c("Strasse", "Hausnummer", "Hausnummernzusatz"),
+   fuzzy_threshold = .9, seed = 1234)
> head(df2_matched$ret)

```

	Var1	Var2	Strasse	Hausnummer	Hausnummernzusatz	STRNAME	STRSL
2897	3	B	Aalbeckestr.	7	<NA>	AALBECKESTRAßE	76567
1467	3	A	Abboweg	6	<NA>	ABBOWEG	70002
2589	1	A	Adalbertstr.	149	<NA>	ADALBERTSTRAßE	70009
12	1	B	Adelenstr.	12	<NA>	ADELENSTRAßE	70011
1063	4	C	Adlerstr.	40	A	ADLERSTRAßE	70013
1562	3	B	Adlerstr.	50	<NA>	ADLERSTRAßE	70013

	HNR	HNRZ	RW	HW	UBZ
2897	7	<NA>	399331.5	5709651	324
1467	6	<NA>	387004.9	5708323	843
2589	149	<NA>	390150.8	5706988	032
12	12	<NA>	397968.0	5705219	432
1063	40	A	392083.5	5707898	022
1562	50	<NA>	391972.3	5707893	022

```
> table(df2_matched$QA$qAddress)
```

1	2	3
2950	49	1

49 records had invalid house numbers and one record was matched by fuzzy matching. This record can be inspected in detail.

```
> id <- which(df2_matched$QA$qAddress == 3)
> df2_matched$ret[id,]
```

Var1	Var2	Strasse	Hausnummer	Hausnummernzusatz	STRNAME
2238	3	B St.-Georg-Str.	10	<NA>	SANKT-GEORG-STRAßE
STRSL	HNR	HNRZ	RW	HW	UBZ
2238	72670	10	<NA>	396545	5706070 531

```
> df2_matched$QA[id,]
```

qAddress	qscore
3000	3 0.8470588

In this case the fuzzy matching procedure was most likely correct (St.-Georg-Str. matched SANKT-GEORG-STRAßE).

The number of cases with correct street name and randomly assigned house numbers is 10.

```
> sum(df2_matched$QA$qscore == 0)
```

```
[1] 10
```