Explaining the dataset: differences between TwiNL and the Twitter API

Erik Tjong Kim Sang

Netherlands eScience Center

Twi-XL, Hilversum, 21 November 2022

netherlands eScience center

What is this?

The TwiNL data set is a large collection of tweets written in Dutch

We created this collection to give researchers and students easier access to Dutch tweets

Next to data access, we also wanted to make it easier for researchers and students to analyze the tweets: by making summaries of results and data visualizations

netherlands eScience center

History

- Since December 2010 we have collected tweets written in Dutch
- The collection currently contains 4.5 billion Dutch tweets (2010-2022)
- It contains tweet text and tweet meta data but no other media (images)
- The data format is JSON (gzipped)
- The project TwiNL offered analysis of this data: twiqs.nl (2012-2019)
- The data are available for usage in the Twi-XL project
- Organizations with a similar data set: RuG and SURF

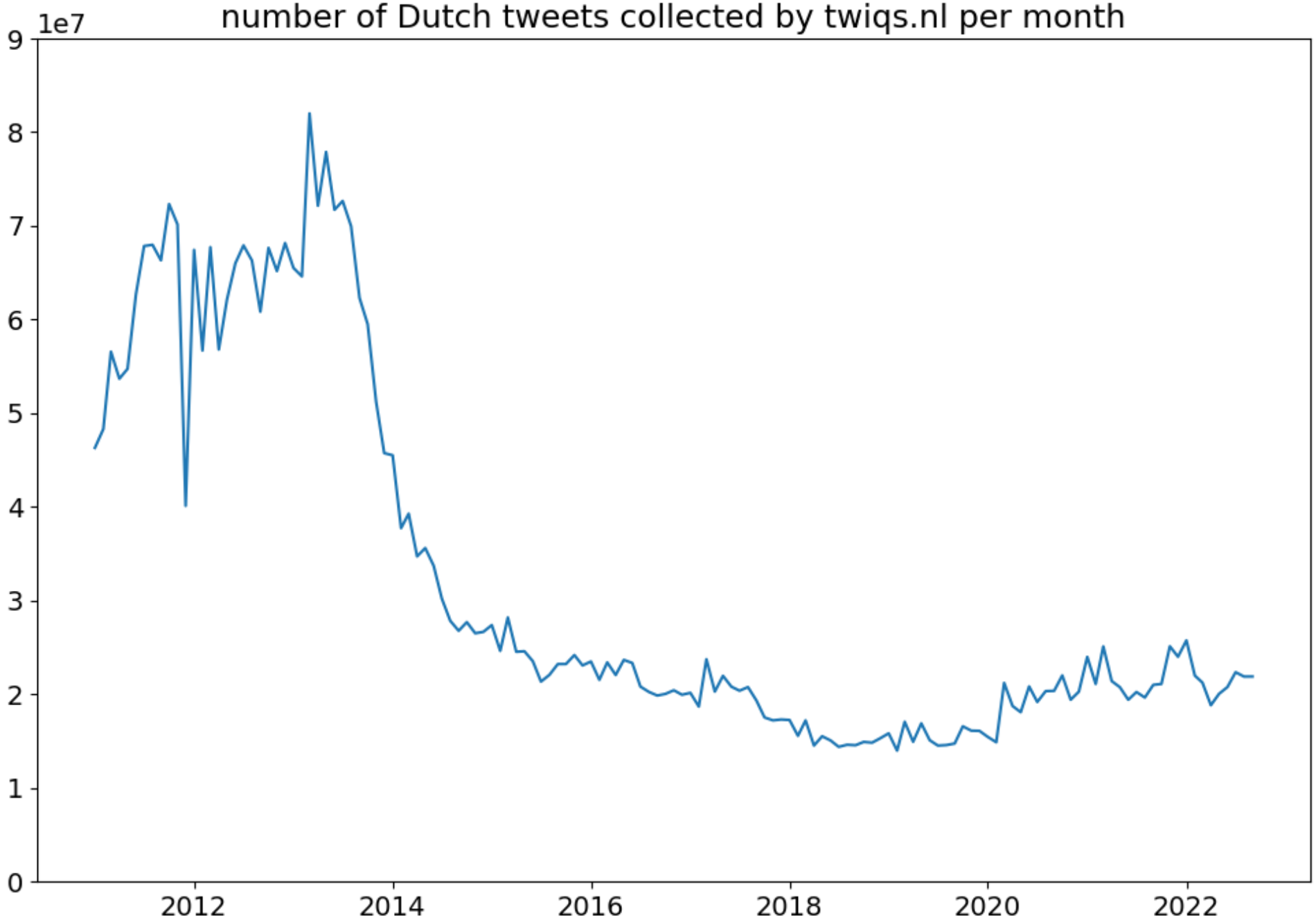
netherlands eScience center

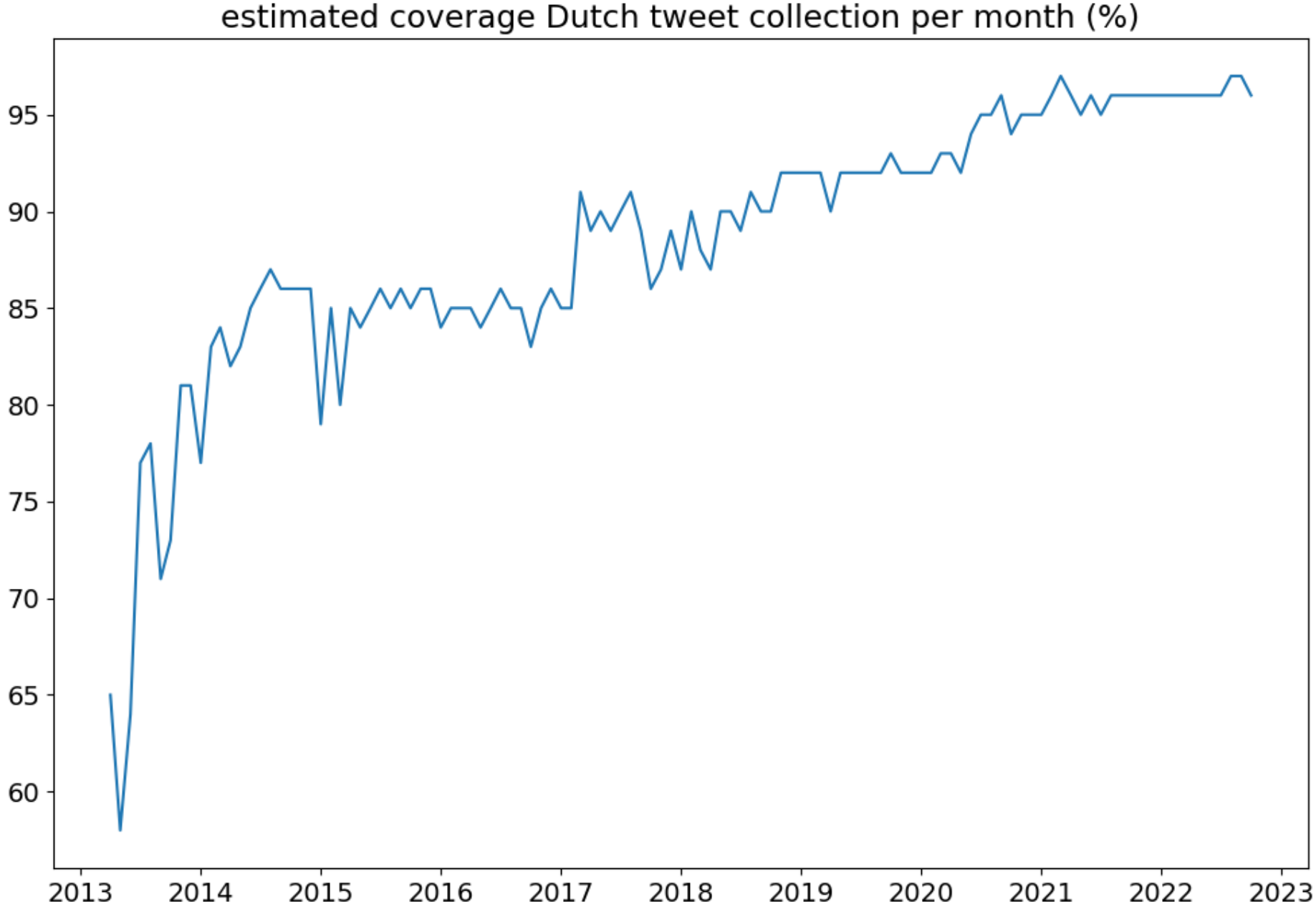
Collecting the data

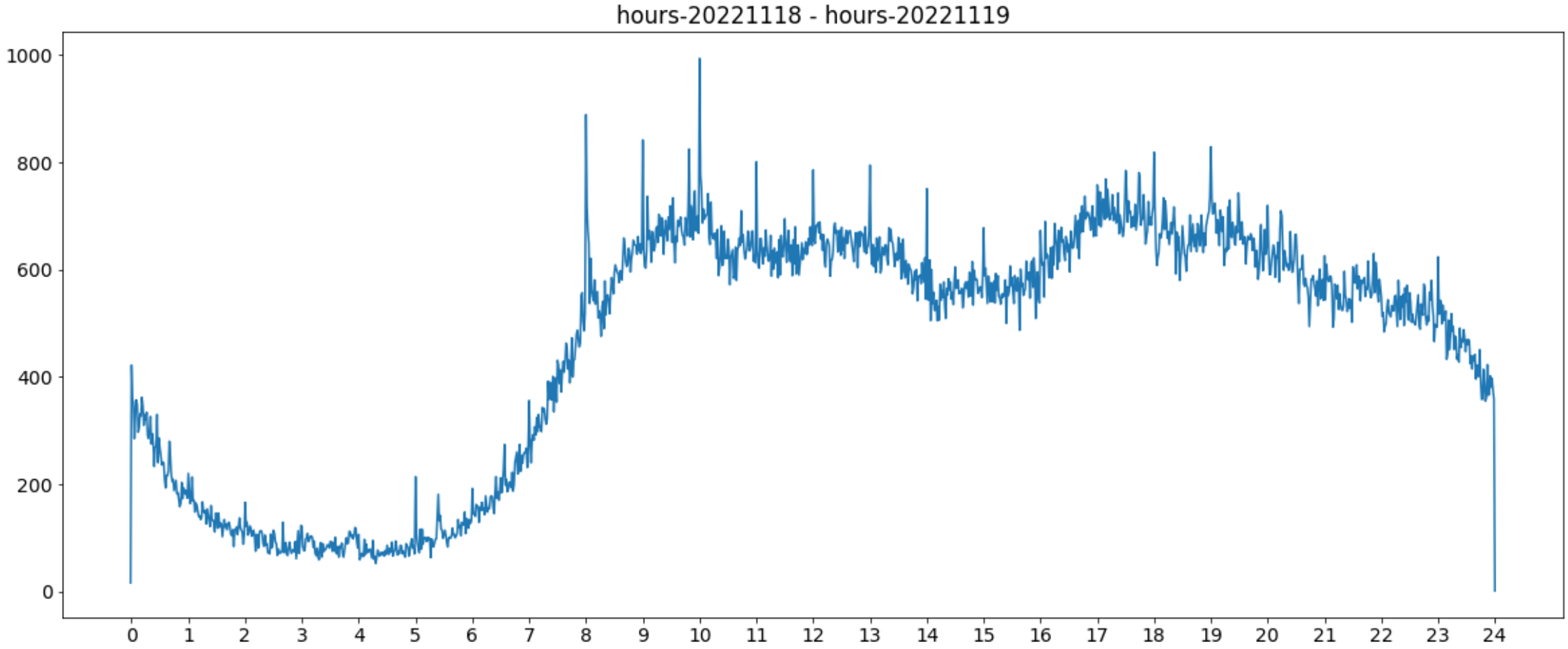
We use four strategies for selecting tweets from Twitter:

1. 800 frequent Dutch Twitter terms from the year 2020
2. 5,000 most frequent authors of Dutch tweets
3. Geographical location: NL, Flanders, Dutch Antilles (6), Suriname
4. 160 dialect words: Flemish, street language, Limburgish, …

The collection also contains foreign tweets: **always filter on: lang: "nl"**


number of Dutch tweets collected by twiqs.nl per month
1e7
9
8
7
6
5
4
3
2
1
0
2012
2014
2016
2018
2020
2022


estimated coverage Dutch tweet collection per month (%)
95
90
85
80
75
70
65
60
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023


hours-20221118 - hours-20221119
1000
800
600
400
200
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

netherlands eScience center

Applications of the tweet collection

- Dutch dialect research
- Election result prediction
- Various linguistic research, for example emerging new words
- Opinion mining related to events
- Opinion mining related to tv programs
- Opinion mining related to anti-COVID measures
- Various research related to language usage and users on Twitter

netherlands eScience center

Limitations of the tweet collection

- We cannot give you the tweets of the collection, only their ids
- If you zoom in on specific dates and specific topics, not many tweets might be left
- The authors of tweets are not representative for the Dutch population

See: **Predicting the 2011 Dutch Senate Election Results with Twitter**, by Erik Tjong Kim Sang and Johan Bos. Proceedings of SASN 2012, the EACL 2012 Workshop on Semantic Analysis in Social Networks, ACL, Avignon, France, 2012, pages 53-60, https://ifarm.nl/erikt/papers/sasn2012.pdf

netherlands eScience center

More information

Dealing with Big Data: the Case of Twitter, by Erik Tjong Kim Sang and Antal van den Bosch. In: Computational Linguistics in the Netherlands Journal, volume 3, ISSN: 2211-4009, pages 121-134, 2013, https://ifarm.nl/erikt/papers/clin2013.pdf

Other work with the data set: see papers that cite this paper: https://scholar.google.nl/scholar?cites=17473617881621327219

The former website associated with the collection: http://twiqs.nl