unibz
Freie Universität Bozen
Libera Università di Bolzano
Università Liedia de Bulsan

Converting html tables into a Stata dataset

Jan Ditzen

Free University of Bozen-Bolzano, Bozen, Italy

www.jan.ditzen.net, jan.ditzen@unibz.it

https://janditzen.github.io/htmltab2stata/

May 9, 2024

Motivation

- Data often available online in form of html Tables.
- Copying content of tables to Stata slow and manual.
- `htmltab2stata` automatises downloading html Tables directly into a Stata dataset.
- Restricted to HTML "<table>" environment.

htmltab2stata

Syntax

`htmltab2stata , url(string) [tablenumber(string) firstrow href]`

- `url(url)` the url of the html website to be processed. The url has to be a downloadable html website. url can be a web address or a local html file.
- `tablenumber(integer)` number of table within the html document. Default is 1, i.e. the first table is processed.
- `firstrow` Use firstrow of table as variable names.
- `href` Links enclosed in are added to the content transferred to Stata.

Example

Tables on My Github Page

A hillwalking economist working on panel time series with a love for coding econometric methods.

Edit profile

92 followers · 23 following

Free University of Bolzano-Bozen
www.jan.ditzen.net
@janditzen
https://bsky.app/profile/janditzen.bsky.social
in/jan-ditzen-54584477

Achievements

Beta Send feedback

JanDitzen / README.md

Welcome

I am an Assistant Professor (RTD-A) at the Free University of Bozen-Bolzano, Bolzano, Italy.

My research interests are in the field of Applied Econometrics with a focus on Panel-Time Series and Spatial Econometrics and Growth Empirics. I have also an interest in Growth Theory, Simulation Studies and the implementation of econometric methods into statistical software.

See also my Personal webpage.

Stata packages - Econometric Methods

Package	Version	Updated	Description	Article/Slides
xtdcce2	release v4.6	release date january	Estimation of Common Correlated Effects (CCE) Estimator. Includes xtcd2 and xtcse2 to test for and estimate exponent of cross-section dependence	Stata Journal 18:3, Stata Journal 21:3, Slides.
xtbreak	release v1.5	release date april	Estimating and testing for many known and unknown structural breaks in time series and panel data. Work with Yiannis Karavias and Joakim Westerlund.	arXiv 2110.14550, arXiv 2211.06707 (Main paper), Slides.
nwxtregress	release v0.3	release date xxxxx xxx 2023	Network Regressions in Stata with unbalanced panel data and time varying network structures or spatial weight matrices. Work with William Grieser and Morad Zekhnini.	Slides
xthst	release v1	release date june 2020	Testing for slope homogeneity in Stata. Work with Tore Bersvendsen.	Stata Journal 21:1
xtgrangert	release v1.14	release date august 2023	Improved tests for Granger noncausality in panel data. Work with Jiaqi Xiao, Yiannis Karavias, Artūras Juodis and Vasilis Sarafidis.	Stata Journal 23:1, Slides
xtnumfac	release v1.2	release date xxxxx xxx 2023	A battery of estimators for the number of common factors in time series and panel-data models. Work with Simon Reese.	Stata Journal 23:2.

Stata packages - Data Processing

Package	Version	Updated	Description
simulate2	release v.109	release date october 2023	Enhanced and parallised simulations in Stata.
mmat2tex	release v2.2	release date december 2023	Export Mata Matrix to LaTeX Tables.
htmltab2stata	release v1.1	release date august 2023	Converting html tables into a Stata dataset.
stataid	release v1	release date january 2019	Obtaining and displaying information about running Stata instances and closing Stata instances in Microsoft Windows.
multishell	release v.1	release date july 2022	Parallise loops in Stata (discontinued, no further bug fixing/ development).

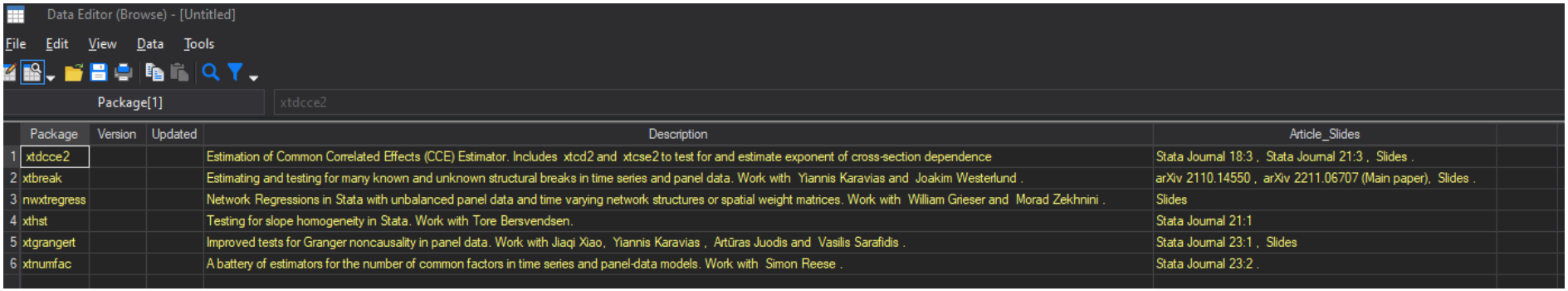
Example

Tables on My Github Page

- There are two tables on My GitHub page. The pages are defined using the HTML "<table>" environment.
- Let's say we want to load the first table.

```
htmltab2stata , url(https://github.com/JanDitzen) firstrow href
```

- Downloads the website, parses it for tables and loads the first table into a new Stata dataset:



	Package	Version	Updated	Description	Article_Slides
1	xtdcce2			Estimation of Common Correlated Effects (CCE) Estimator. Includes xtcd2 and xtcse2 to test for and estimate exponent of cross-section dependence	Stata Journal 18:3 , Stata Journal 21:3 , Slides .
2	xtbreak			Estimating and testing for many known and unknown structural breaks in time series and panel data. Work with Yiannis Karavias and Joakim Westerlund .	arXiv 2110.14550 , arXiv 2211.06707 (Main paper), Slides .
3	nwxtregress			Network Regressions in Stata with unbalanced panel data and time varying network structures or spatial weight matrices. Work with William Grieser and Morad Zekhnini .	Slides
4	xthst			Testing for slope homogeneity in Stata. Work with Tore Bersvendsen.	Stata Journal 21:1
5	xtgrangert			Improved tests for Granger noncausality in panel data. Work with Jiaqi Xiao, Yiannis Karavias , Artūras Juodis and Vasilis Sarafidis .	Stata Journal 23:1 , Slides
6	xtnumfac			A battery of estimators for the number of common factors in time series and panel-data models. Work with Simon Reese .	Stata Journal 23:2 .

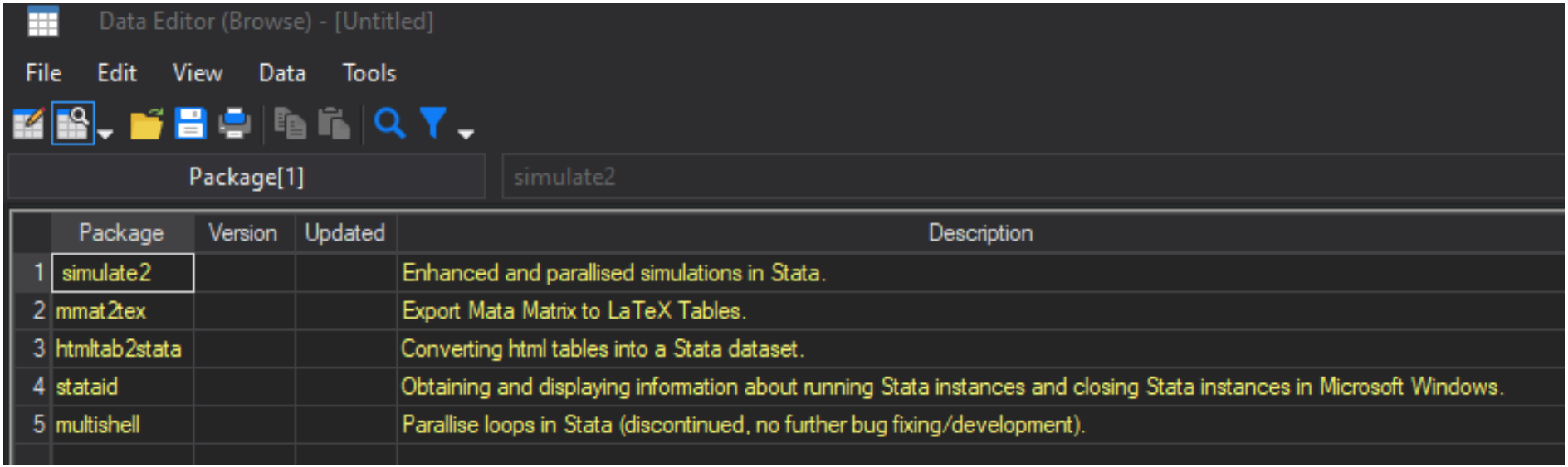
Example

Tables on My Github Page

- Let's say we want to load the second table.

```
htmltab2stata , url(https://github.com/JanDitzen) firstrow href tablenumber(2)
```

- Leads to:



	Package	Version	Updated	Description
1	simulate2			Enhanced and parallised simulations in Stata.
2	mmat2tex			Export Mata Matrix to LaTeX Tables.
3	htmltab2stata			Converting html tables into a Stata dataset.
4	stataid			Obtaining and displaying information about running Stata instances and closing Stata instances in Microsoft Windows.
5	multishell			Parallise loops in Stata (discontinued, no further bug fixing/development).

Conclusion

- `hmtltab2stata` is a simple tool to load html Tables into Stata datasets.
- Can load websites from URLs or files.
- Can processes multiple Tables per page and load directly links.
- Limited to static pages with HTML "<table>" environment.
- How to install?

```
net install htmltab2stata , from("https://janditzen.github.io/htmltab2stata/")
```

- More info:

jan.ditzen.net

GitHub