nsgtm:

National Statistics: Guatemala

Using R to access official data from the Guatemalan National Institute of Statistics

Oscar de León
Universidad del Valle de Guatemala
@uRos2018
September 14, 2018

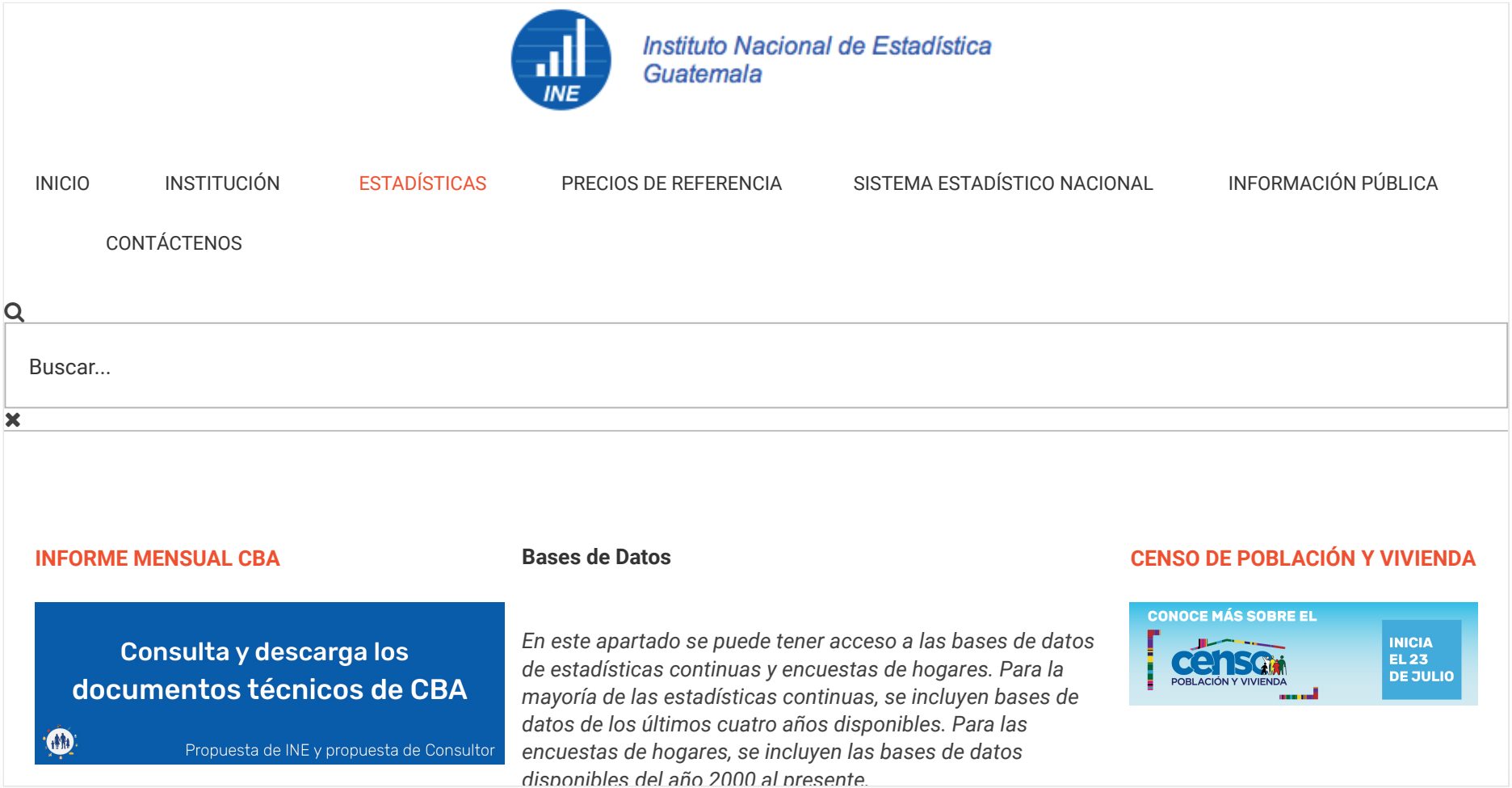
Access to official statistics

Other available tools

Access to official statistics (GSBPM 7.4) (from this Awesome list)

- R package rsdmx. Easy access to data from statistical organisations that support SDMX webservices. The package contains a list of SDMX access points of various national and international statistical institutes.
- R package oecd Search and Extract Data from the OECD
- R package sorvi Finnish Open Government Data Toolkit
- R package eurostat Tools to download data from the Eurostat database together with search and manipulation utilities.
- R package acs Download, Manipulate, and Present American Community Survey and Decennial Data from the US Census.
- R package inegiR Access to data published by INEGI, Mexico's official statistics agency.
- R package cbsodataR. Access to Statistics Netherlands' (CBS) open data API from R.

Web portal for the Guatemalan NSI



The `nsgtm` package

A catalog of datasets from the NSI

```
(topics <-   nsgtm::topics_gtm   %>%
   sample_n(3))
```

category	period	year	section	text	resource_uri
175	1	2012	Publicaciones	Mayo 2012	https://www.ine.gob.gt/sistema/uploads/2014/02/24/FLE47r3w
152	1	2015	Boleta / Diccionario de Variables	Diccionario de variables	https://www.ine.gob.gt/sistema/uploads/2017/01/16/emO14nj4
80	1	2015	Estadisticas de Transportes y Servicios / Estadísticas de Turismo	I - Flujo Migratorio Año 2015	https://www.ine.gob.gt/sistema/uploads/2016/12/14/fflzORJDQ

Check for file availability

Some resources, although listed in the portal, can sometimes be taken down.

```
(    topics  %>%
  select(text, year, resource_uri) %>%
  mutate(
    # Get the resource availability from the NSI portal
    data = map(
      resource_uri,
      nsgtm::check_resource
    )
  ) %>%
  tidyr::unnest() %>%
  select(text, year, file_available, file_size, file_units, size_readable) )
```

text	year	file_available	file_size	file_units	size_readable
Mayo 2012	2012	TRUE	64074	bytes	62.6 KiB
Diccionario de variables	2015	TRUE	422492	bytes	412.6 KiB

And you have direct access to download the file and read the data

(Now using all the file types found in the catalog)

```
( resources <-    nsgtm::topics_gtm  %>%
    mutate(type = tools::file_ext(resource_uri)) %>%
    left_join(
      nsgtm::resources_gtm
    ) %>%
    group_by(type) %>%                          # Get the smallest file of each type
    slice(which.min(file_size)) )
```

category	period	year	section	text	resource_uri
136	1	2013	3.11. Indicadores de violencia intrafamiliar	Número de denuncias de violencia intrafamiliar	https://www.ine.gob.gt/sistema/uploads/2014/07/24/VW7r
175	1	2014	Publicaciones	Publicación	https://www.ine.gob.gt/sistema/uploads/2014/10/07/9xxell

You can proceed to get the data

```
                                                #-------------------------------------*
resources %>%                                   # starting with the selected resources
  split(.$resource_uri) %>%                     # for each one:
  walk(                                         #-------------------------------------*
    ~ {
      cat(                                      # start a new slide (with ##)
        "\n##",
        gsub("^[[:space:]]*", "", .x$text),     # showing the dataset name
        " (file type: ", .x$type, ")",          # and the file type
        " {.smaller}\n\n",
        sep = ""
      )
      resource <- .x$resource_uri               # use the provided uri
      nsgtm::get_resource(resource) %>%         # attempt to read the file
      head() %>%                                # and show the top 6 rows
      knitr::kable() %>%
      print()
    }
  )
```

Principales resultados INACIF exhumaciones (file type: xls)

Servicios médico de exhumaciones por mes, según departamento1/	X__1	X__2	X__3	X__4	X__5	X__6	X__7	X__8	X__9	X__10	X__11	X__12	X__13
Año 2012	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Departamento	Total	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
Total país	174	18	23	18	6	16	12	10	22	13	13	8	15
Guatemala	34	1	3	3	1	1	4	1	7	6	4	0	3
El Progreso	8	0	1	0	0	0	1	2	3	0	0	0	1
Sacatepéquez	2	0	0	0	0	0	0	0	0	0	0	2	0

INACIF exhumaciones (file type: sav)

num_corre	mes_ocu	depto_ocu
1	1	13
2	1	19
3	1	19
4	1	9
5	1	11
6	1	16

Número de denuncias de violencia intrafamiliar (file type: csv)

Año	República	Guatemala	El Progreso	Sacatepéquez	Chimaltenango	Escuintla	Santa Rosa	Sololá	Totonicapán	Quetzaltenango	Suchitepéque
2008	23721	5117	978	926	1167	1417	495	537	616	1435	129
2009	31497	6692	1108	1368	1688	1654	1051	838	698	1442	166
2010	32017	5792	1101	1356	1690	1163	953	994	686	1661	172
2011	33484	6122	1101	1492	1920	1088	857	882	627	1628	195
2012	36107	6911	1039	1651	1785	1169	1082	1106	581	1706	226
2013	36170	6581	1142	1783	1723	794	1441	1062	571	2057	248

Publicación Agosto 2014 (file type: jpg)

```
## Warning in nsgtm::get_resource(resource):
## File type should be one of csv, sav, xls, or xlsx.
## Attempted to read type "jpg".
## Returned an empty dataset.
```

Boleta (file type: pdf)

```
## Warning in nsgtm::get_resource(resource):
## File type should be one of csv, sav, xls, or xlsx.
## Attempted to read type "pdf".
## Returned an empty dataset.
```

INACIF exhumaciones (file type: xlsx)

no_corre	año_ocu	dia_ocu	dia_sem_ocu	mes_ocu	depto_ocu
1	2016	20	4	1	13
2	2016	25	2	1	11
3	2016	28	5	1	13
4	2016	12	6	2	11
5	2016	18	5	2	12
6	2016	23	3	2	19

Use case

Landing page for a "single dataset"

-Base de Datos:

Defunciones	Abrir SPSS
Defunciones fetales	Abrir SPSS
Divorcios	Abrir SPSS
Matrimonios	Abrir SPSS
Nacimientos	Abrir SPSS

-Diccionario de Variables:

Certificado / informe de defunción	Abrir pdf
Diccionario de variables	Abrir xls

Check for file availability

Some resources, although listed in the portal, can sometimes be taken down.

```
# Using the catalog available in the package
(birth_availability <-    nsgtm::topics_gtm  %>%
  # Focus on a specific subset of data
  filter(
    grepl("nacimien", text, ignore.case = TRUE),
    section == "Base de Datos"
  ) %>%
  select(text, year, resource_uri) %>%
  mutate(
    # Get the resource availability from the NSI portal
    data = map(
      resource_uri,
      nsgtm::check_resource
    )
  ) %>%
    tidyr::unnest())
```

Download files

```
# Get data for the first listed dataset
(birth_availability %>%
  slice(1) %>%
  pull(resource_uri) %>%
  nsgtm::get_resource() %>%
  head()
 )
```

Depreg	mupreg	Mesreg	Añoreg	Depocu	Mupocu	Areag	Libras	Onzas	Diaocu	Mesocu	Añoocu	Sex
1	0101	8	9	1	0101	1	5	2	28	6	9	
1	0101	3	10	1	0101	1	5	6	2	4	9	
1	0101	2	9	1	0101	1	5	10	30	1	9	
1	0101	1	10	1	0101	1	6	0	7	9	9	
1	0101	7	9	1	0101	1	6	10	10	7	9	

Further development necessary

More tools for the analyst using R

- Shorthand functions for common tasks
  - filtering topics or resources
  - exporting datasets to a new file in an arbitrary format
  - more file types
- Documentation
- Interface to update the catalog if desired / necessary

Other options for people who don't use R

- Also include facilities to export each dataset in a format of their choice (this needs some R backend, e.g. `shiny`)

Other steps to cover

- Hunt down dictionaries and data manuals
- Similar tool/package for geodata (different official sources)
- Other sources of official data:

Guatemalan science council Open data portal

Oscar de León
Uninversidad del Valle de Guatemala
odeleon@ces.uvg.edu.gt

Thanks

Dependencies

package	dependencies
haven	forcats, hms, Rcpp, readr, tibble, magrittr, rlang, methods, pkgconfig, utils, R6, cli, crayon, pillar, assertthat, grDevices, fansi, utf8, tools
readr	Rcpp, tibble, hms, R6, methods, pkgconfig, rlang, utils, cli, crayon, pillar, assertthat, grDevices, fansi, utf8, tools
readxl	cellranger, Rcpp, tibble, rematch, methods, utils, cli, crayon, pillar, rlang, assertthat, grDevices, fansi, utf8, tools
tibble	cli, crayon, methods, pillar, rlang, utils, assertthat, grDevices, fansi, utf8, tools

Unique dependencies:
forcats, hms, Rcpp, readr, tibble, magrittr, rlang, methods, pkgconfig, utils, R6, cli, crayon, pillar, assertthat, grDevices, fansi, utf8, tools, cellranger, rematch (total of 21 dependencies)