R PACKAGES FOR SDG 6.2 DATA

Development and MSc Data Science project

Linda Karani

Lars Schöbitz

AGENDA

1. R Package Development
2. Master Thesis Project
    - Fit models for estimates
        - Prepare primary indicators estimates based on JMP estimation rules
        - Prepare secondary indicator service levels
    - Extract r squared and p-values from models
        - Identify those with low values < 0.20 and high values > 0.80
        - Plots exploring the OLS regression using two countries with low and high rsq values
    - Plots showing the model fit coefficients(r.squared and p value)
    - Fit alternative models and assess goodness of fit
    - Next steps
3. Questions for Discussion

R PACKAGE DEVELOPMENT

R DATA PACKAGE - BENEFITS

- Data accessible as a single table for any analysis tool
- Data can be imported to R using one command
- Public website with detailed documentation _ e.g. `washdata` R Package https://katilingban.io/washdata/index.html

R DATA PACKAGE - SANITATION

- Data in long format (19,528 rows)
- 9 variables

iso3	source	type	year	var_short	var_long	residence	san_service_
AFG	MICS	Survey	2003	s_imp_u	Improved	urban	user interfac
AFG	NRVS	Survey with microdata	2005	s_imp_u	Improved	urban	user interfac
AFG	NVRA	Survey with microdata	2008	s_imp_u	Improved	urban	user interfac
AFG	MICS	Survey with microdata	2011	s_imp_u	Improved	urban	user interfac

R DATA PACKAGE - NEW VARIABLES

- `residence`: urban/rural/national
- `san_service_chain` (Sanitation Service Chain):

san_service_chain	n
open defecation	2774
sharing	1553
user interface	12663
containment	195
emptying	1356
transport	10
FS treatment	85
WW treatment	921
NA	3

R DATA PACKAGE - USE CASES

1. Using JMP methods to reproduce estimates and apply different models (Linda Karani - MSc Data Science)
2. Writing an R Package with a function to produce estimates (and a function to produce service ladder plots)

R DATA PACKAGE - USE CASES

1. Using JMP methods to reproduce estimates and apply different models (Linda Karani - MSc Data Science)
2. Writing an R Package with a function to produce estimates (and a function to produce service ladder plots)

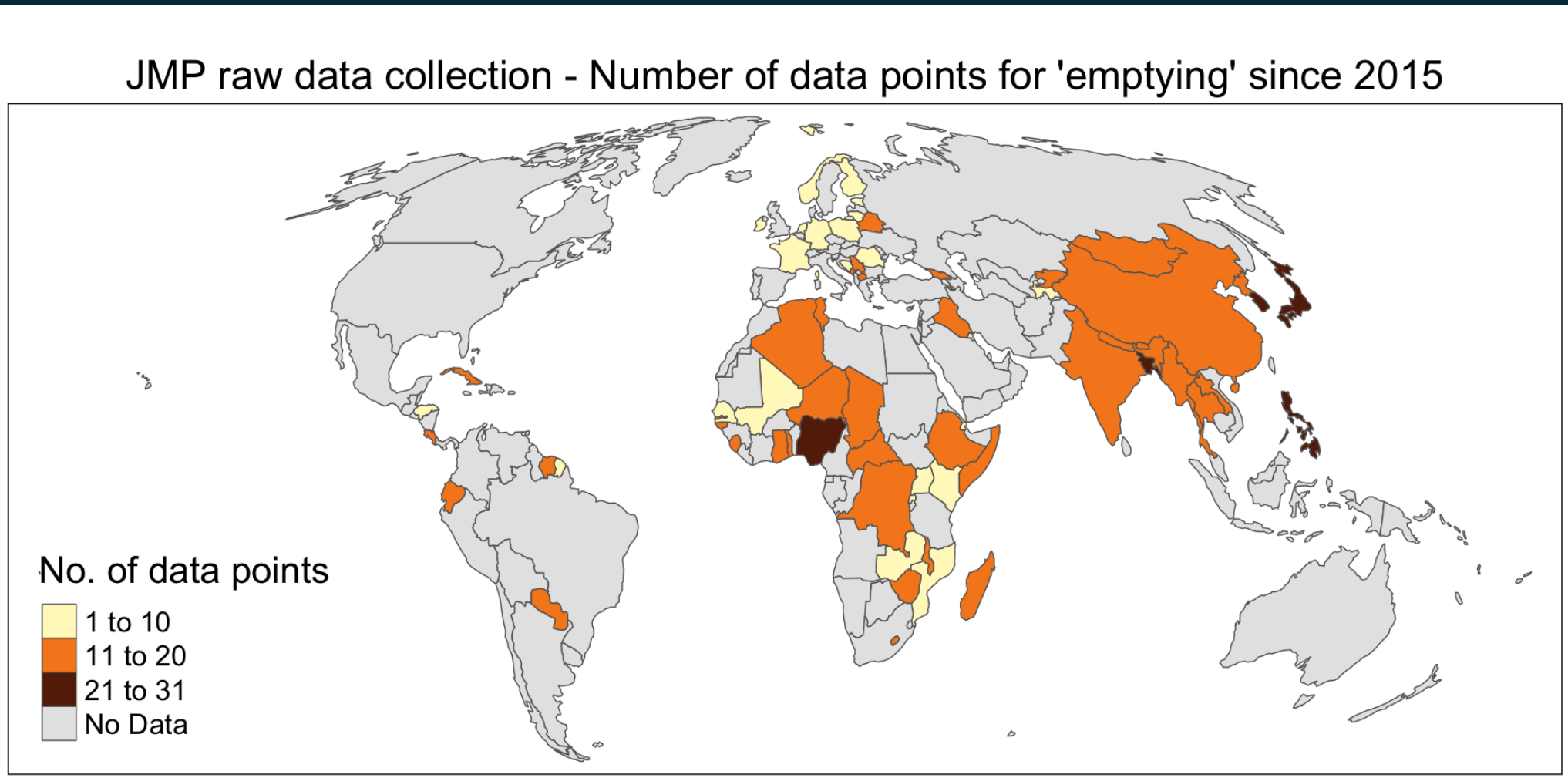
```r
estimate(iso3 = "AFG",           # default: all iso3 codes
         year = 2010:2030,       # Single year or range of years
         var_short = NULL,       # default: all variables (NULL)
         residence = "national") # default: national
```

R DATA PACKAGE - USE CASES

1. Using JMP methods to reproduce estimates and apply different models (Linda Karani - MSc Data Science)
2. Writing an R Package with a function to produce estimates (and a function to produce service ladder plots)

```r
estimate(iso3 = "AFG",           # default: all iso3 codes
         year = 2010:2030,       # Single year or range of years
         var_short = NULL,       # default: all variables (NULL)
         residence = "national") # default: national
```

3. Great potential for unforeseen use cases enabled by making the data readily accessible (research, teaching, joining with other data, etc.)


JMP raw data collection - Number of data points for 'emptying' since 2015
No. of data points
1 to 10
11 to 20
21 to 31
No Data

JMP raw data collection: Number of data points for ‘emptying’ since 2015

country	n
Philippines	62
Nigeria	48
Bangladesh	40
Japan	40
South Korea	32
Ethiopia	20
Niger	20
Belarus	16
China	16
Congo - Kinshasa	16

R DATA PACKAGE - WHAT'S NEXT?

- Submission of proposal for further development to ORD (Open Research Data) Programme of ETH Domain (15k in-kind + 15k ETH Board), due: 12th December
- Submission of a proposal to Colorado WASH Symposium (focus on Linda's work), due: 25th November

https://ethrat.ch/wp-content/uploads/2022/10/Application-guidelines_TrackC-Contribute-projects-3rd-call-1.pdf

MASTER THESIS PROJECT

MASTERS THESIS OUTLINE

OBJECTIVES

- Generate sanitation estimates for rural and urban from raw packaged data using documented methods
  - only the primary indicators and derived secondary service levels:
    - basic sanitation service
    - limited sanitation service
    - unimproved sanitation service
    - no sanitation service
- Fit different statistical models to assess goodness of fit

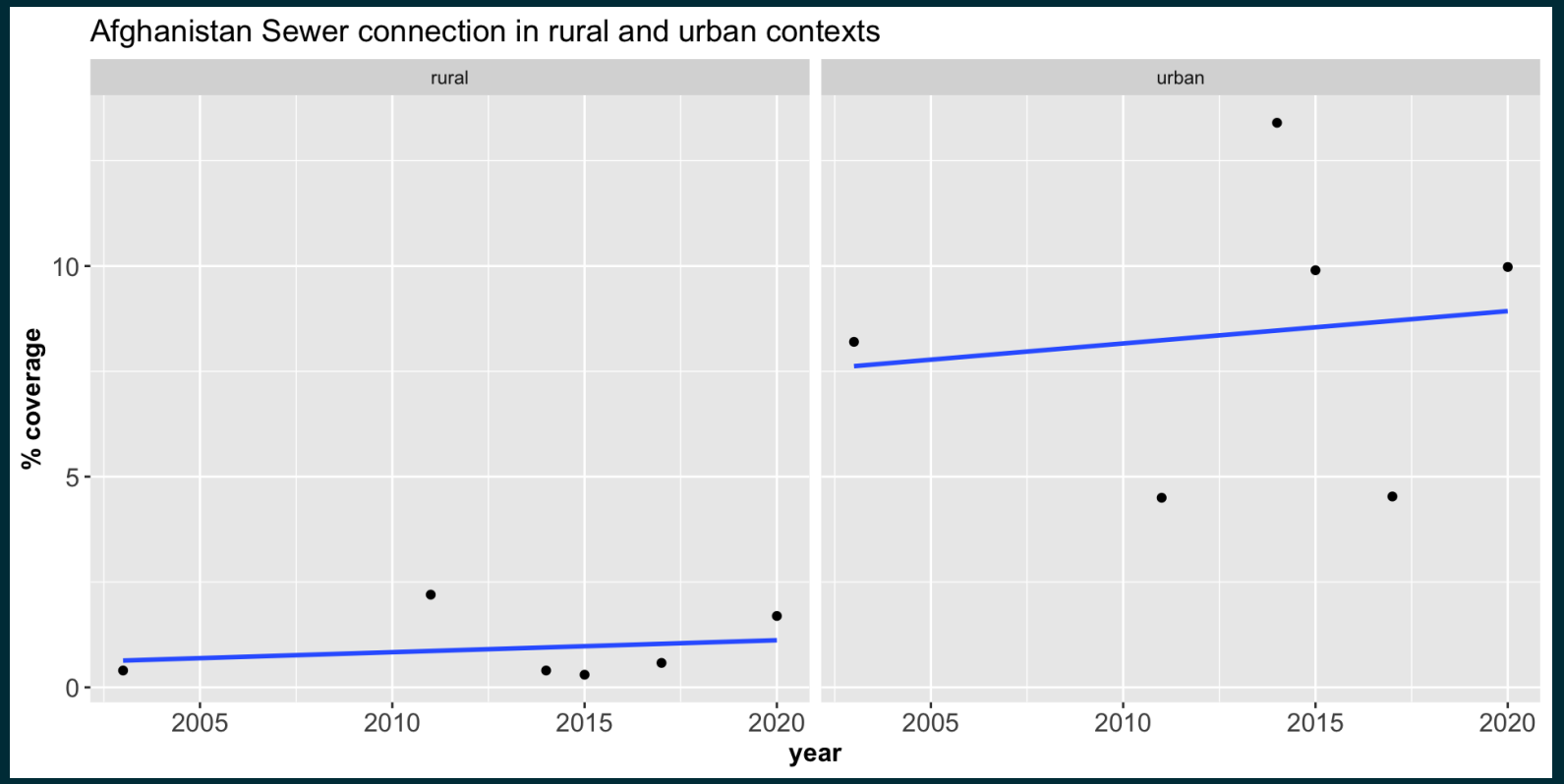
STATUS

- Managed to derive estimates for the primary indicators and derived secondary service levels

PLOT LINEAR FIT FOR DIFFERENT COUNTRIES

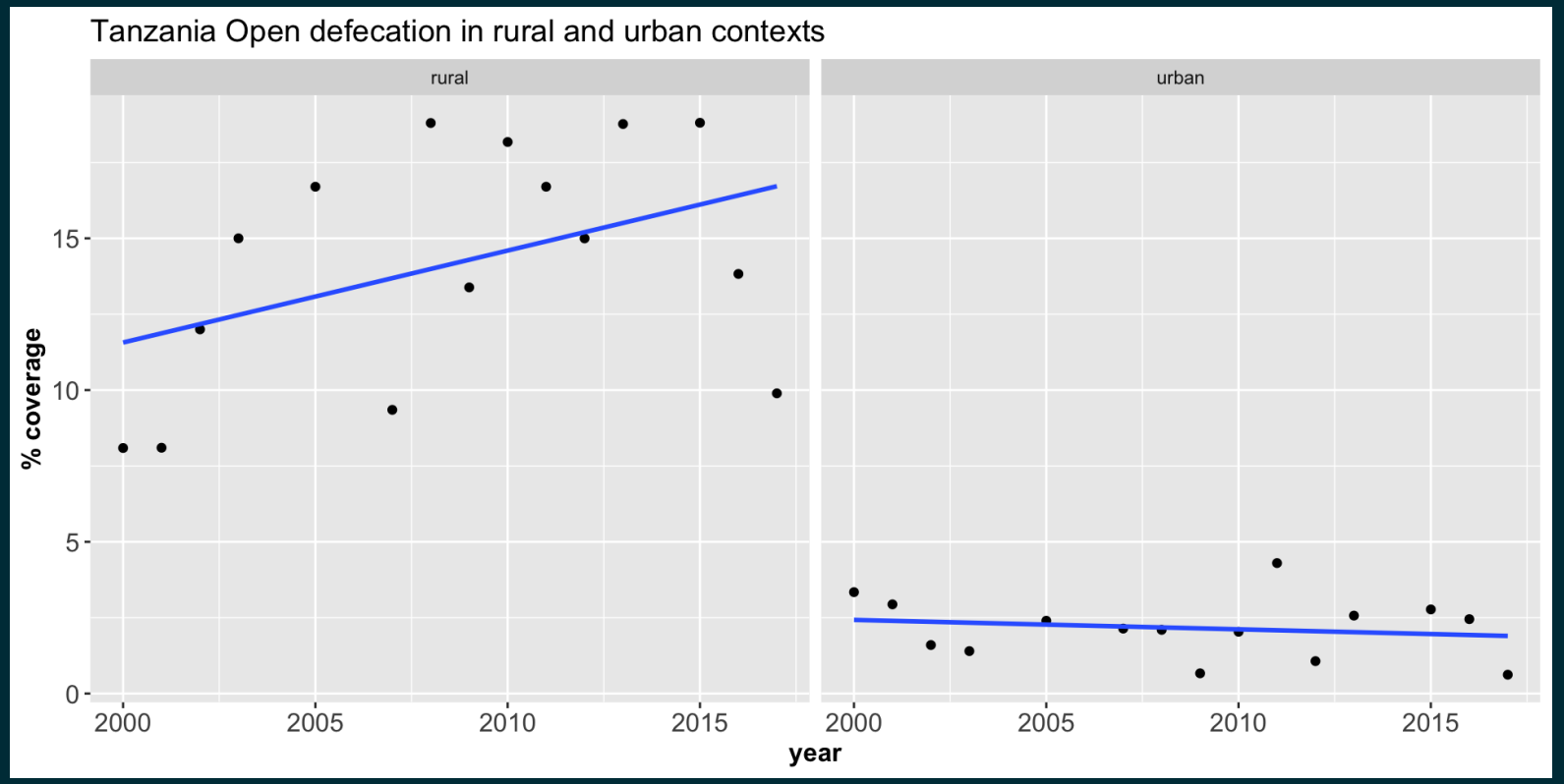
OLS REGRESSION : R SQUARED < 0.20

AFGHANISTAN SEWER CONNECTION



- Afghanistan r.sq rural = 0.0426018
- Afghanistan r.sq urban = 0.0170736

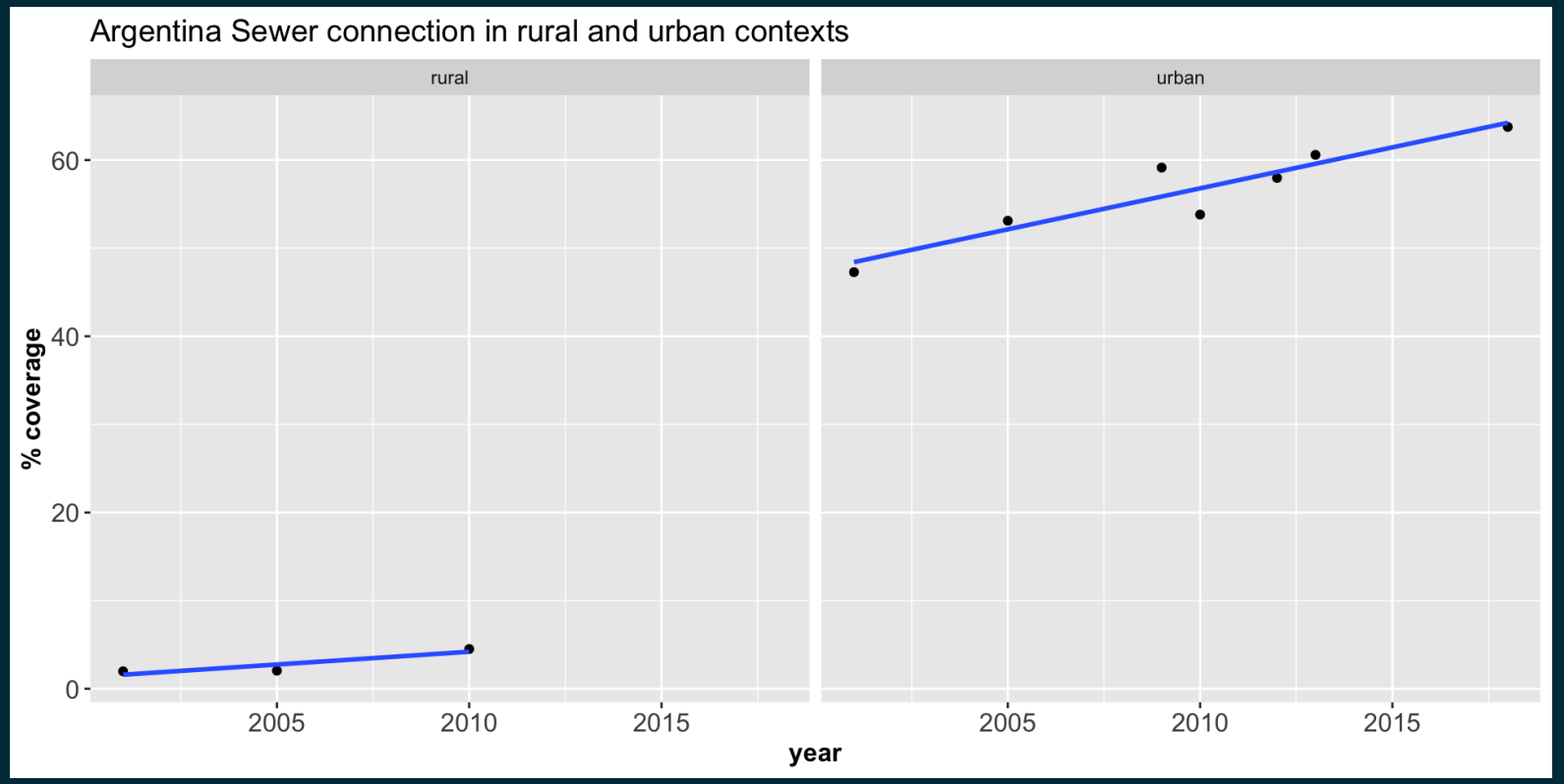
TANZANIA OPEN DEFECATION



- Tanzania r.sq rural = 0.1824534
- Tanzania r.sq urban = 0.0296412

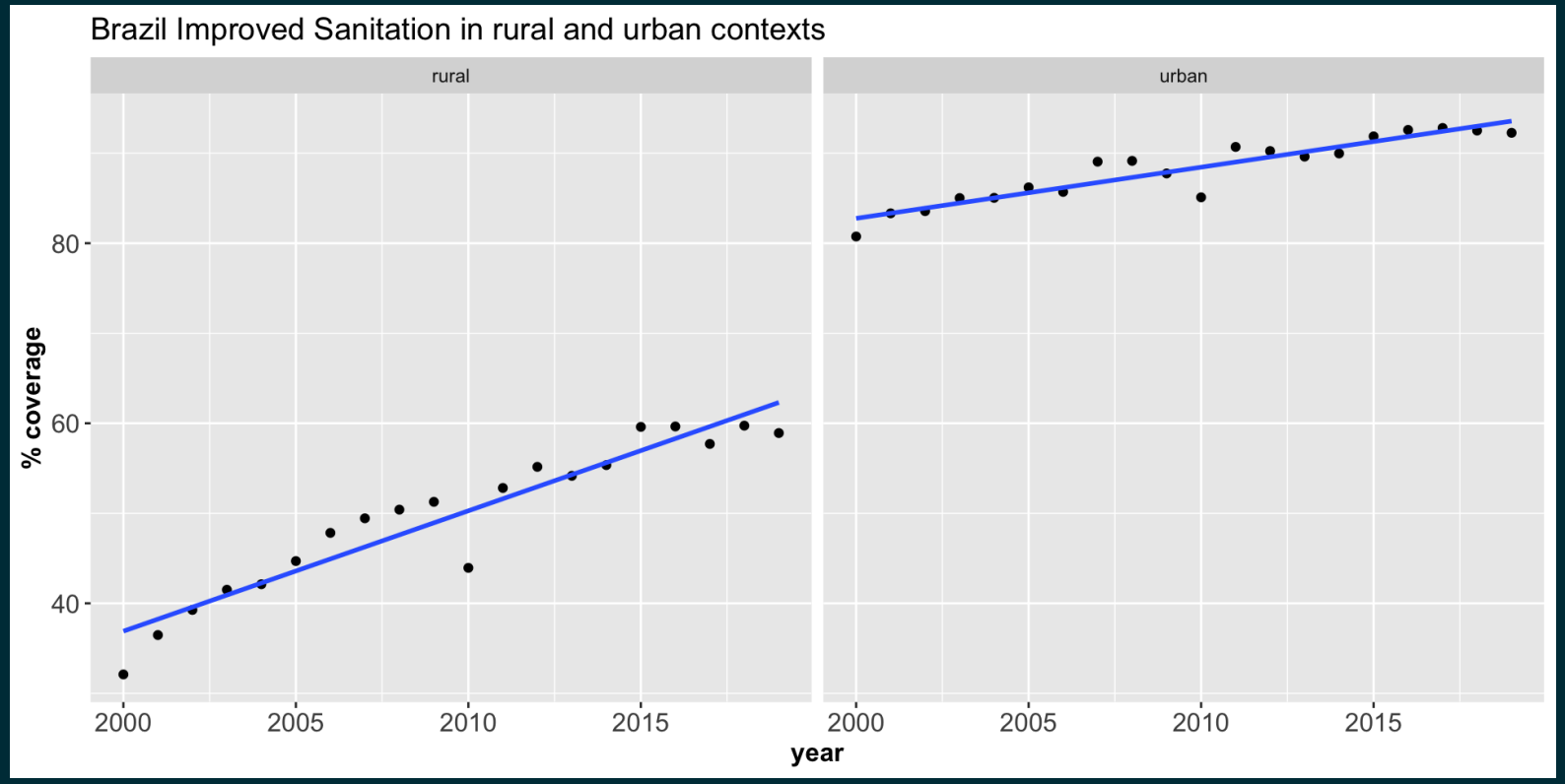
OLS REGRESSION : R SQUARED > 0.80

ARGENTINA SEWER CONNECTION OLS REGRESSION



- Argentina r.sq rural = 0.82
- Argentina r.sq urban = 0.87

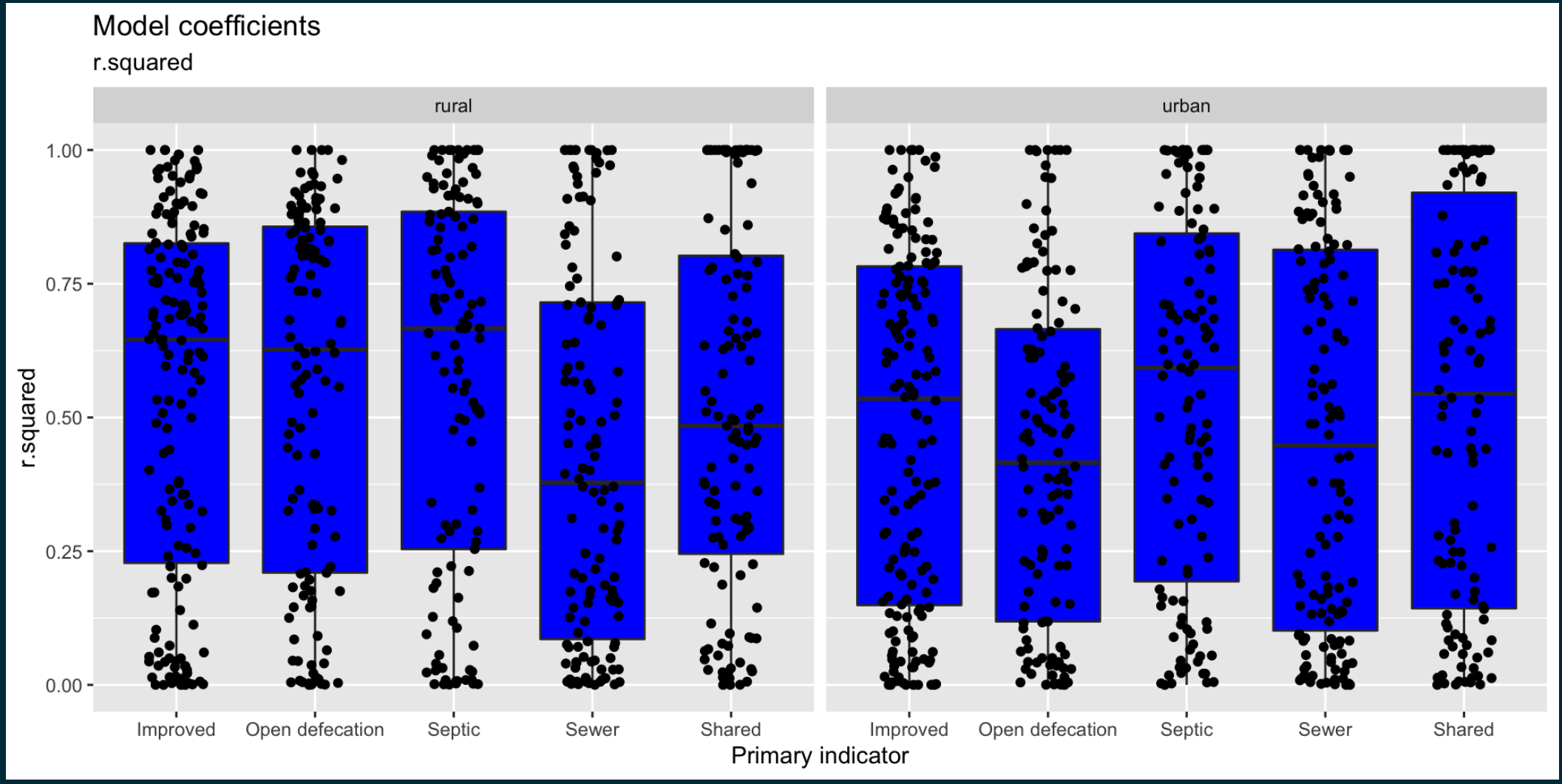
BRAZIL IMPROVED SANITATION OLS REGRESSION



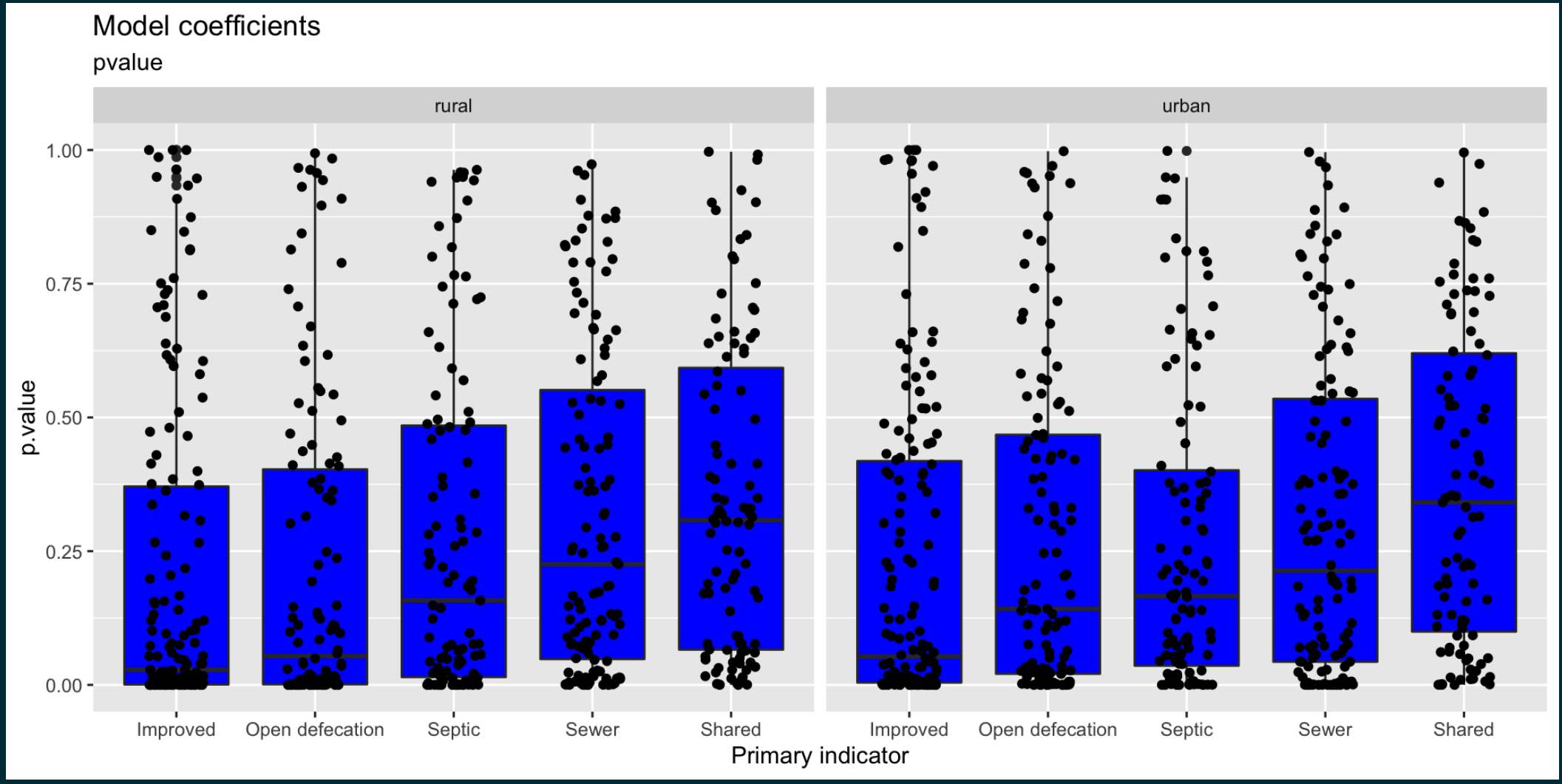
- Brazil r.sq rural = 0.90
- Brazil r.sq urban = 0.87

PLOTTING MODEL COEFFICIENTS

MODEL FIT : R SQUARED



MODEL FIT : P VALUE

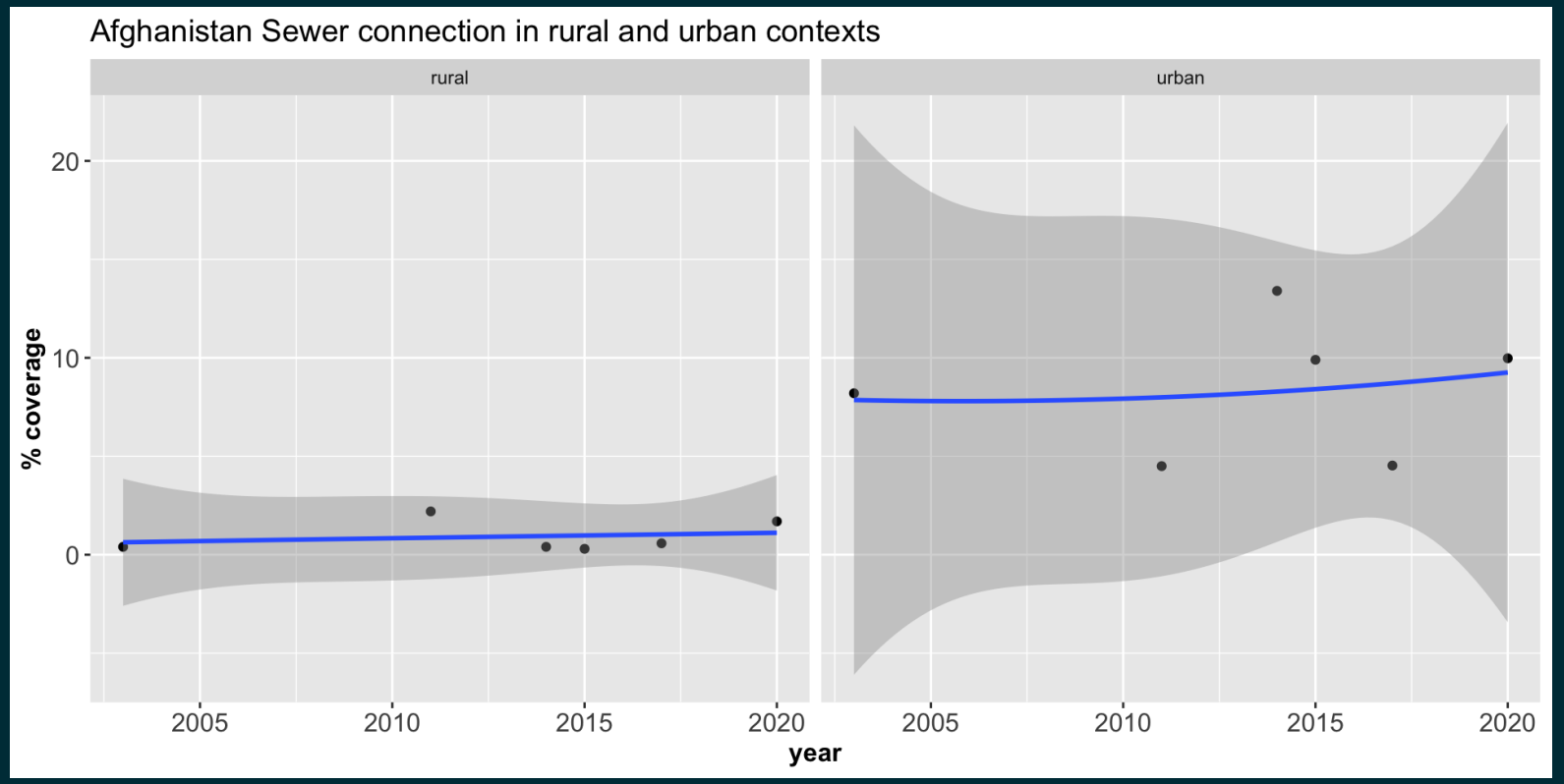


ALTERNATIVE MODELS

- Fitting different models to the data to compare goodness of fit

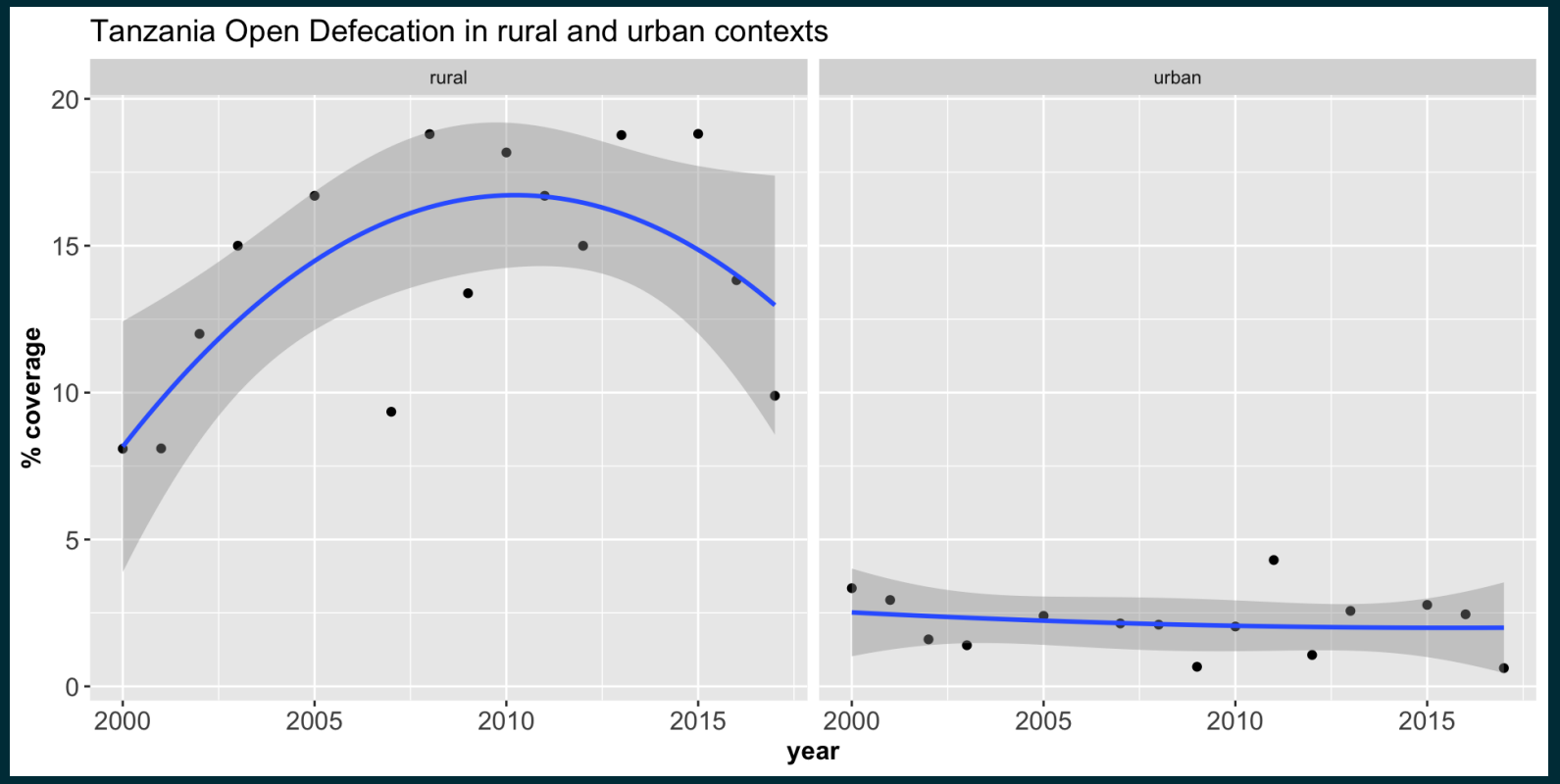
PLOT USING A 4TH ORDER POLYNOMIAL

AFGHANISTAN SEWER CONNECTION



- Afghanistan r.sq rural = 0.043
- Afghanistan r.sq urban = 0.022

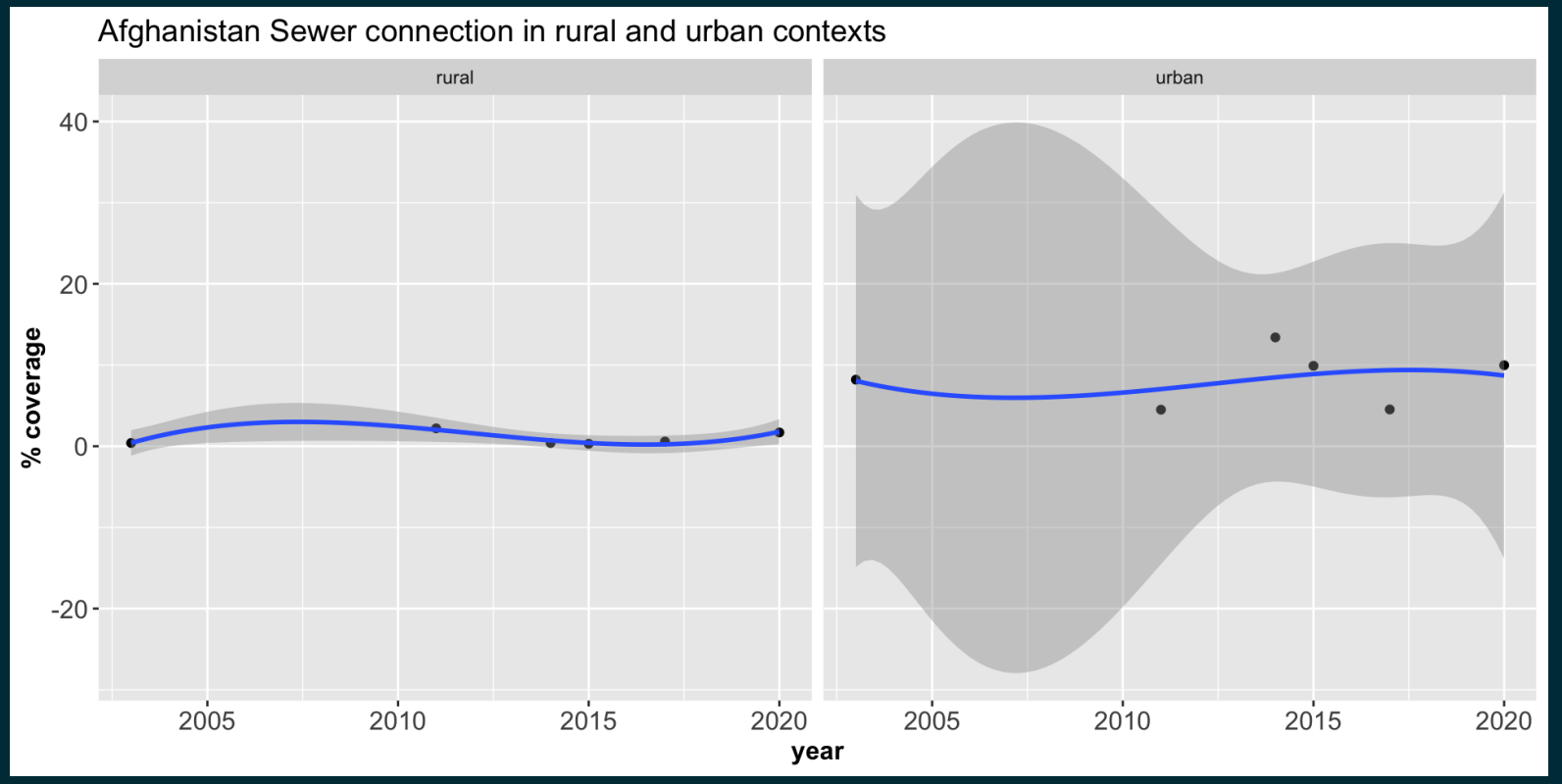
TANZANIA OPEN DEFECATION



- Tanzania r.sq rural = 0.487
- Tanzania r.sq urban = 0.033

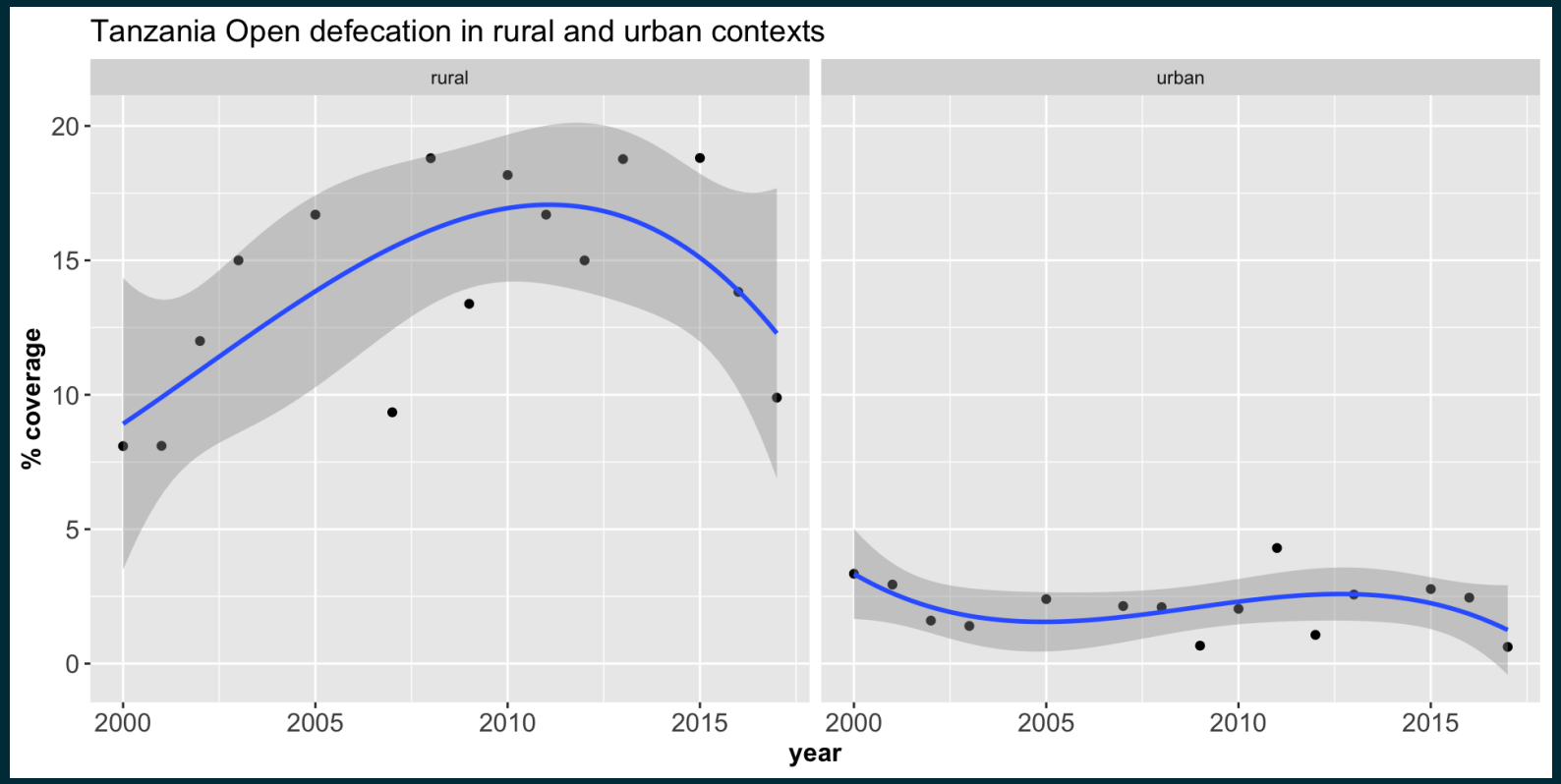
PLOT USING SPLINES

AFGHANISTAN SEWER CONNECTION



- Afghanistan r.sq rural = 1.00
- Afghanistan r.sq urban = 1.00

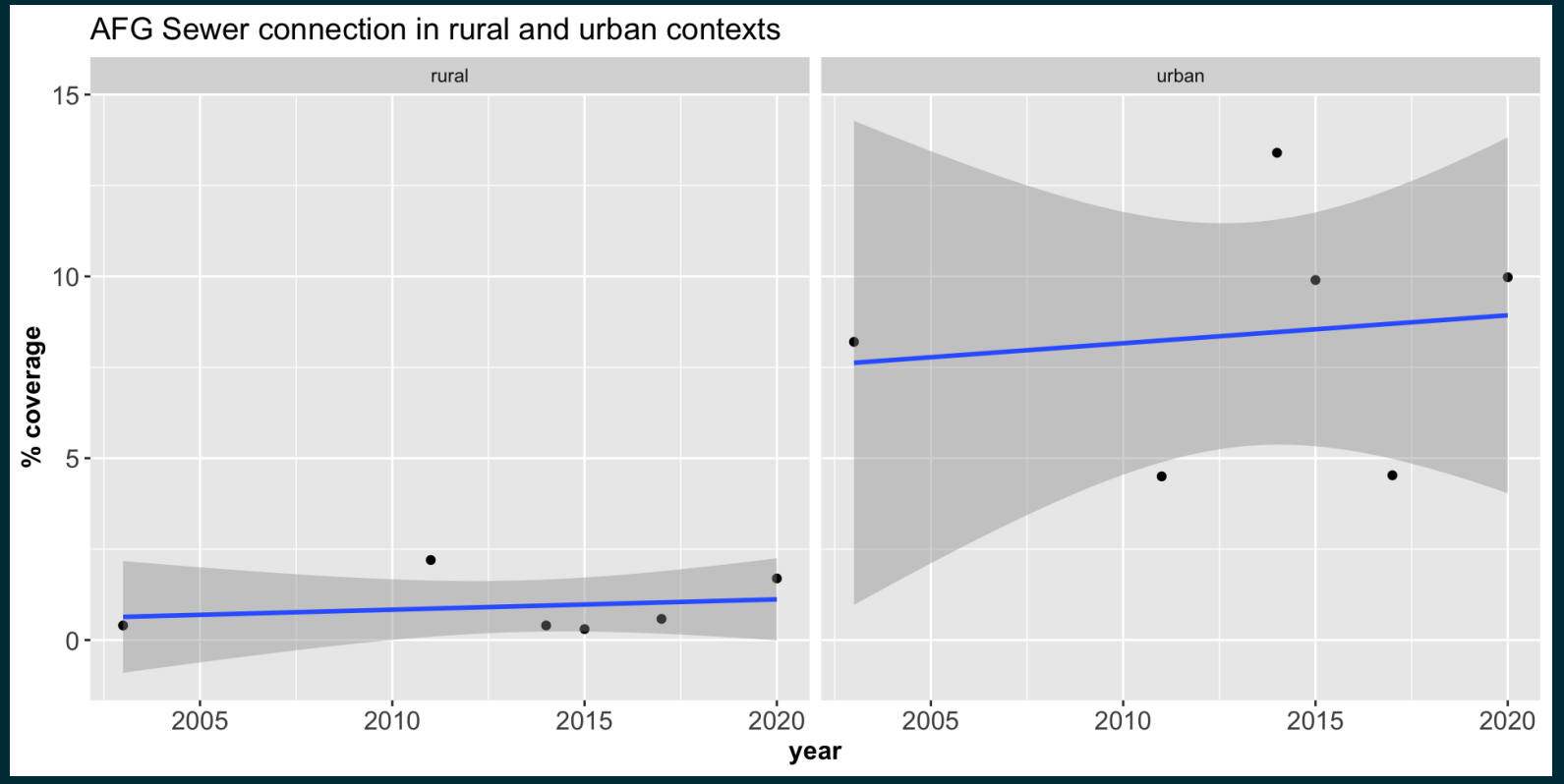
TANZANIA OPEN DEFECATION



- Tanzania r.sq rural = 0.65
- Tanzania r.sq urban = 0.37

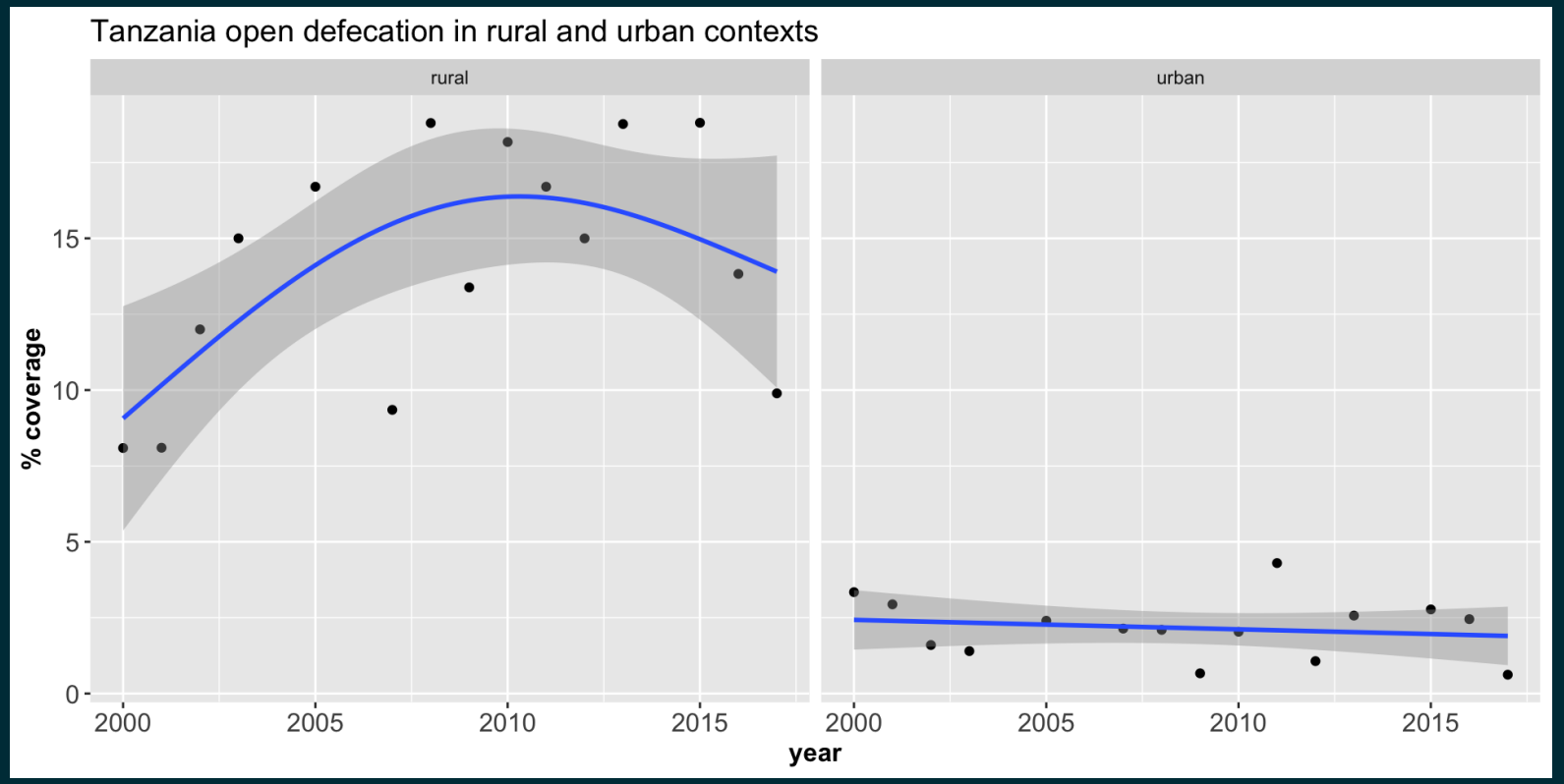
PLOT USING GAM

AFGHANISTAN SEWER CONNECTION



- Afghanistan BIC rural = 18.5
- Afghanistan BIC urban = 36.1

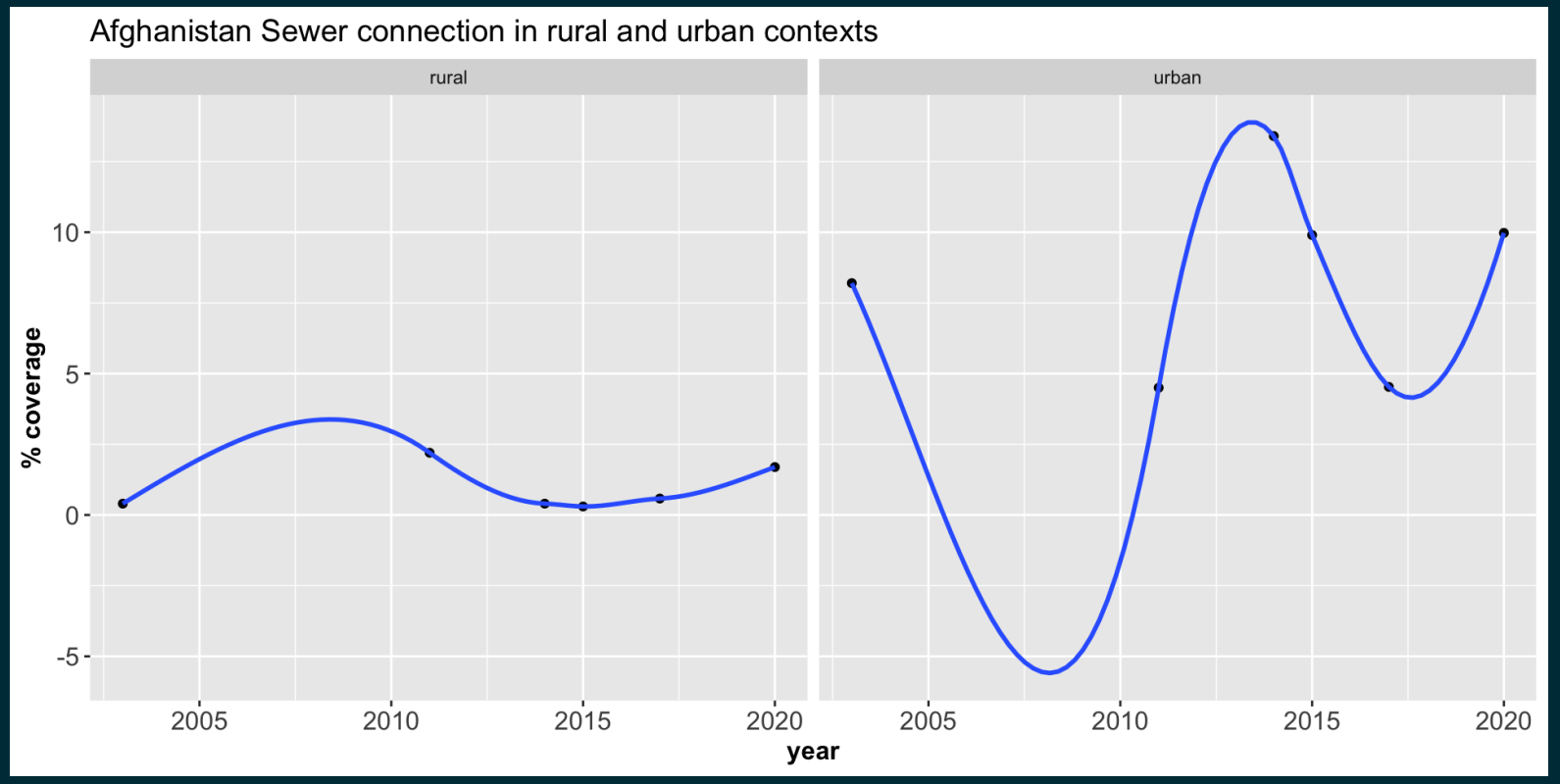
TANZANIA OPEN DEFECATION



- Tanzania BIC rural = 87.6
- Tanzania BIC urban = 49.2

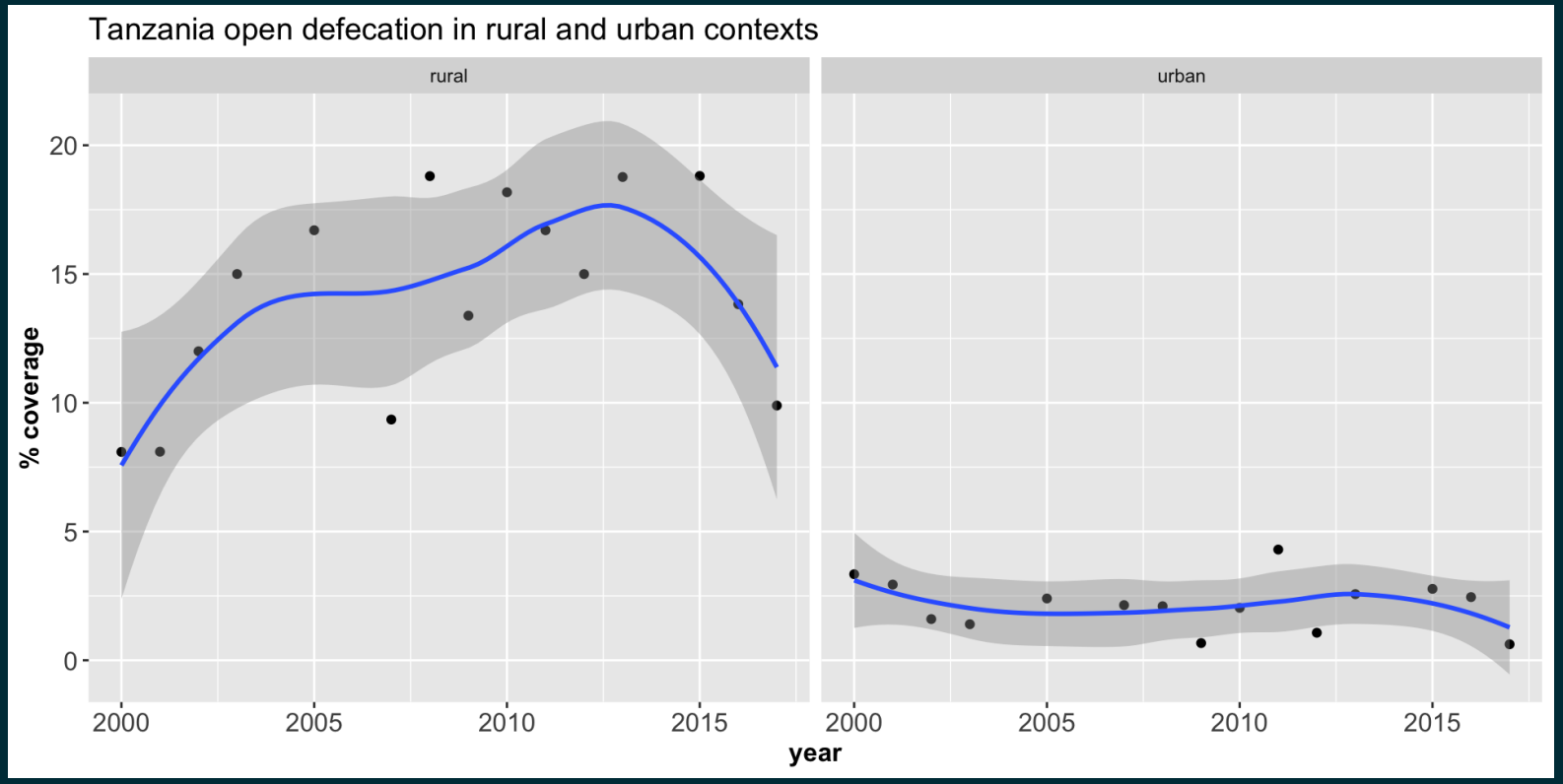
PLOTTING USING LOESS

AFGHANISTAN SEWER CONNECTION



- Afghanistan RSE rural = 0.043
- Afghanistan RSE urban = 0.022

TANZANIA OPEN DEFECATION



- Tanzania RSE rural = 0.487
- Tanzania RSE urban = 0.033

SUMMARY

COMPARISON OF MODELS

1. Higher order polynomial
    - Pros
        - Easy implementation
        - Significant increase in r.sq for above 10 data points and for coverage values above 10%. The r.sq for open defecation in Tanzania increased from 18% to 49%
    - Cons
        - r.sq does not seem to improve for coverage values below 10% and for few data points
2. Splines
    - Pros
        - Takes the shape of the data
        - r.sq significantly improves. 100% for Afghanistan sewer connection in rural and urban contexts
    - C

- Cons
    - Selection of knots

3. Generalized additive models
    - Pros
        - Automatic selection of knots
    - Cons
        - Tends to have a large standard error
        - Not easily explainable

4. Loess
    - Pros
        - Takes the shape of the data
        - Low standard error for few data points
    - Cons
        - Tends to overfit for few data points
        - Not easily explainable
        - Not suitable for prediction

NEXT STEPS

- Work through secondary indicators
- Create a decision tree that recommends different models for different countries

DISCUSSION

QUESTIONS

1. How are country files updated on the server?
    - How often? All together, or one by one as there is new data?
2. The current database for raw data shows 379 different sources. They are all abbreviated. Do you have a table where all these abbreviations are spelled out?
3. Ratio RS1

> “Other ratios used for ‘basic’ indicators (RW1, RS1) are calculated using simple averages” - JMP Methodology, March 2018

THANKS

THANKS!

Data source: washdata.org

Slides - source code: https://github.com/KaraniLinda/SDG-6-2-Reproducibility/blob/main/slides/r-packages-for-sdg62-data-lkarani-lschoebitz.qmd

Slides - PDF download: https://github.com/KaraniLinda/SDG-6-2-Reproducibility/raw/main/slides/r-packages-for-sdg62-data-lkarani-lschoebitz.pdf