Introduction to Geostatistics

introductory statistics for earth scientists

Edzer Pebesma

edzer.pebesma@uni-muenster.de
Institute for Geoinformatics (**ifgi**)
University of Münster

summer semester 2007/8,
April 14, 2010

Course practicalities

Language vorlesungen: German; Exercises C: English (German), rest: German

Test English + German, Multiple Choice

Books Wonnacott & Wonnacott: Introductory Statistics (WW)

Symbols will follow WW, with AS coloured

Exercises will use R, `http://www.r-project.org/`

ifgi

Course practicalities

Language	vorlesungen: German; Exercises C: English (German), rest: German
Test	English + German, Multiple Choice
Books	Wonnacott & Wonnacott: Introductory Statistics (WW)
Symbols	will follow WW, with AS coloured
Exercises	will use R, `http://www.r-project.org/`

Course practicalities

Language vorlesungen: German; Exercises C: English (German), rest: German

Test English + German, Multiple Choice

Books Wonnacott & Wonnacott: Introductory Statistics (WW)

Symbols will follow WW, with AS coloured

Exercises will use R, `http://www.r-project.org/`

Course practicalities

Language	vorlesungen: German; Exercises C: English (German), rest: German
Test	English + German, Multiple Choice
Books	Wonnacott & Wonnacott: Introductory Statistics (WW)
Symbols	will follow WW, with AS coloured
Exercises	will use R, `http://www.r-project.org/`

Course practicalities

Language	vorlesungen: German; Exercises C: English (German), rest: German
Test	English + German, Multiple Choice
Books	Wonnacott & Wonnacott: Introductory Statistics (WW)
Symbols	will follow WW, with AS coloured
Exercises	will use R, `http://www.r-project.org/`

Groups

- A1 Mi 10-12 Uhr, CIP B, Robert-Koch-Str., Lydia Gerharz, Christoph Stasch (Ben Gräler)
- A2 Mi 10-12 Uhr, CIP C, Robert-Koch-Str., Kristina Helle, Tutorin: Merret Buurman
- A3 Mi 10-12 Uhr, CIP RvE, Weseler Str., Juliane Exeler, Tutor: Kolja Domnick
- B Do 14-16 Uhr, CIP A, Robert-Koch-Str., Katharina Henneböhl, Tutor: Christian Höppner
- C (English), Wed 12-14 Uhr, CIP RvE/ifgi, Edzer Pebesma

Groups

- A1 Mi 10-12 Uhr, CIP B, Robert-Koch-Str., Lydia Gerharz, Christoph Stasch (Ben Gräler)
- A2 Mi 10-12 Uhr, CIP C, Robert-Koch-Str., Kristina Helle, Tutorin: Merret Buurman
- A3 Mi 10-12 Uhr, CIP RvE, Weseler Str., Juliane Exeler, Tutor: Kolja Domnick
- B Do 14-16 Uhr, CIP A, Robert-Koch-Str., Katharina Henneböhl, Tutor: Christian Höppner
- C (English), Wed 12-14 Uhr, CIP RvE/ifgi, Edzer Pebesma

Groups

- A1 Mi 10-12 Uhr, CIP B, Robert-Koch-Str., Lydia Gerharz, Christoph Stasch (Ben Gräler)
- A2 Mi 10-12 Uhr, CIP C, Robert-Koch-Str., Kristina Helle, Tutorin: Merret Buurman
- A3 Mi 10-12 Uhr, CIP RvE, Weseler Str., Juliane Exeler, Tutor: Kolja Domnick
- B Do 14-16 Uhr, CIP A, Robert-Koch-Str., Katharina Henneböhl, Tutor: Christian Höppner
- C (English), Wed 12-14 Uhr, CIP RvE/ifgi, Edzer Pebesma

Groups

- A1 Mi 10-12 Uhr, CIP B, Robert-Koch-Str., Lydia Gerharz, Christoph Stasch (Ben Gräler)
- A2 Mi 10-12 Uhr, CIP C, Robert-Koch-Str., Kristina Helle, Tutorin: Merret Buurman
- A3 Mi 10-12 Uhr, CIP RvE, Weseler Str., Juliane Exeler, Tutor: Kolja Domnick
- B Do 14-16 Uhr, CIP A, Robert-Koch-Str., Katharina Henneböhl, Tutor: Christian Höppner
- C (English), Wed 12-14 Uhr, CIP RvE/ifgi, Edzer Pebesma

Groups

- A1 Mi 10-12 Uhr, CIP B, Robert-Koch-Str., Lydia Gerharz, Christoph Stasch (Ben Gräler)
- A2 Mi 10-12 Uhr, CIP C, Robert-Koch-Str., Kristina Helle, Tutorin: Merret Buurman
- A3 Mi 10-12 Uhr, CIP RvE, Weseler Str., Juliane Exeler, Tutor: Kolja Domnick
- B Do 14-16 Uhr, CIP A, Robert-Koch-Str., Katharina Henneböhl, Tutor: Christian Höppner
- C (English), Wed 12-14 Uhr, CIP RvE/ifgi, Edzer Pebesma

Grading

- Two parts will be graded: exercises and lectures.
- The grade for the lectures is the test result.
- The test will assume understanding of R output (graphs, tables, etc)
- The exercises grade is based on assignment of field work.
- To pass the exercises, you need 50% of hand-ins handed in correctly (these will not be graded).

Grading

- Two parts will be graded: exercises and lectures.
- The grade for the lectures is the test result.
- The test will assume understanding of R output (graphs, tables, etc)
- The exercises grade is based on assignment of field work.
- To pass the exercises, you need 50% of hand-ins handed in correctly (these will not be graded).

Grading

- Two parts will be graded: exercises and lectures.
- The grade for the lectures is the test result.
- The test will assume understanding of R output (graphs, tables, etc)
- The exercises grade is based on assignment of field work.
- To pass the exercises, you need 50% of hand-ins handed in correctly (these will not be graded).

Grading

- Two parts will be graded: exercises and lectures.
- The grade for the lectures is the test result.
- The test will assume understanding of R output (graphs, tables, etc)
- The exercises grade is based on assignment of field work.
- To pass the exercises, you need 50% of hand-ins handed in correctly (these will not be graded).

Grading

- Two parts will be graded: exercises and lectures.
- The grade for the lectures is the test result.
- The test will assume understanding of R output (graphs, tables, etc)
- The exercises grade is based on assignment of field work.
- To pass the exercises, you need 50% of hand-ins handed in correctly (these will not be graded).

Motivation

Geoscientists (such as landscape ecologists, geoinformaticians, human geographers, ...) collect and study geoscientific data, and need to analyse them. How do we do the analysis? first we need the data.

How do we get data? first we need a research question.

What is a good research question?

- one that can be answered by data
- the data needed can be obtained with the research "budget"

Given a good research question, how should we collect data? sampling: what, when, where, how often?

Motivation

Geoscientists (such as landscape ecologists, geoinformaticians, human geographers, ...) collect and study geoscientific data, and need to analyse them. How do we do the analysis? first we need the data.

How do we get data? first we need a research question.

What is a good research question?

- one that can be answered by data
- the data needed can be obtained with the research "budget"

Given a good research question, how should we collect data? sampling: what, when, where, how often?

Motivation

Geoscientists (such as landscape ecologists, geoinformaticians, human geographers, ...) collect and study geoscientific data, and need to analyse them. How do we do the analysis? first we need the data.

How do we get data? first we need a research question.
What is a good research question?

- one that can be answered by data
- the data needed can be obtained with the research "budget"

Given a good research question, how should we collect data?
sampling: what, when, where, how often?

Motivation

Geoscientists (such as landscape ecologists, geoinformaticians, human geographers, ...) collect and study geoscientific data, and need to analyse them. How do we do the analysis? first we need the data.

How do we get data? first we need a research question.
What is a good research question?

- one that can be answered by data
- the data needed can be obtained with the research "budget"

Given a good research question, how should we collect data?
sampling: what, when, where, how often?

Motivation

Geoscientists (such as landscape ecologists, geoinformaticians, human geographers, ...) collect and study geoscientific data, and need to analyse them. How do we do the analysis? first we need the data.

How do we get data? first we need a research question.

What is a good research question?

- one that can be answered by data
- the data needed can be obtained with the research "budget"

Given a good research question, how should we collect data?
sampling: what, when, where, how often?

Motivation

Geoscientists (such as landscape ecologists, geoinformaticians, human geographers, ...) collect and study geoscientific data, and need to analyse them. How do we do the analysis? first we need the data.

How do we get data? first we need a research question.

What is a good research question?

- one that can be answered by data
- the data needed can be obtained with the research "budget"

Given a good research question, how should we collect data? sampling: what, when, where, how often?

Motivation

Geoscientists (such as landscape ecologists, geoinformaticians, human geographers, ...) collect and study geoscientific data, and need to analyse them. How do we do the analysis? first we need the data.

How do we get data? first we need a research question.

What is a good research question?

- one that can be answered by data
- the data needed can be obtained with the research "budget"

Given a good research question, how should we collect data?

sampling: what, when, where, how often?

Motivation

Geoscientists (such as landscape ecologists, geoinformaticians, human geographers, ...) collect and study geoscientific data, and need to analyse them. How do we do the analysis? first we need the data.

How do we get data? first we need a research question.

What is a good research question?

- one that can be answered by data
- the data needed can be obtained with the research "budget"

Given a good research question, how should we collect data?

sampling: what, when, where, how often?

Motivation

Geoscientists (such as landscape ecologists, geoinformaticians, human geographers, ...) collect and study geoscientific data, and need to analyse them. How do we do the analysis? first we need the data.

How do we get data? first we need a research question.

What is a good research question?

- one that can be answered by data
- the data needed can be obtained with the research "budget"

Given a good research question, how should we collect data?

sampling: what, when, where, how often?

Motivation

Geoscientists (such as landscape ecologists, geoinformaticians, human geographers, ...) collect and study geoscientific data, and need to analyse them. How do we do the analysis? first we need the data.

How do we get data? first we need a research question.

What is a good research question?

- one that can be answered by data
- the data needed can be obtained with the research "budget"

Given a good research question, how should we collect data?

sampling: what, when, where, how often?

Induction, deduction

Wikipedia:

Induction: (wp:) is the process of reasoning in which the premises of an argument are believed to support the conclusion but do not entail it; i.e. they do not ensure its truth. Induction is a form of reasoning that makes generalizations based on individual instances. ... Of the candidate systems for an inductive logic, the most influential is Bayesianism. This uses probability theory as the framework for induction. Given new evidence, Bayes' theorem is used to evaluate how much the strength of a belief in a hypothesis should change.

Deduction: is reasoning whose conclusions are intended to necessarily follow its premises. It is more commonly understood as the type of reasoning that proceeds from general principles or premises to derive particulars.

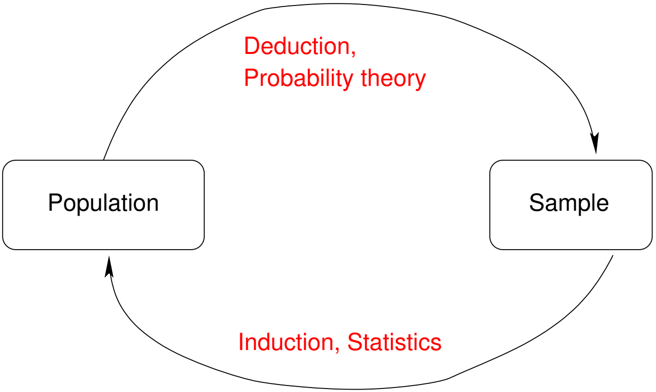
Induction, deduction

Wikipedia:

Induction: (wp:) is the process of reasoning in which the premises of an argument are believed to support the conclusion but do not entail it; i.e. they do not ensure its truth. Induction is a form of reasoning that makes generalizations based on individual instances. ... Of the candidate systems for an inductive logic, the most influential is Bayesianism. This uses probability theory as the framework for induction. Given new evidence, Bayes' theorem is used to evaluate how much the strength of a belief in a hypothesis should change.

Deduction: is reasoning whose conclusions are intended to necessarily follow its premises. It is more commonly understood as the type of reasoning that proceeds from general principles or premises to derive particulars.

Induction, deduction



Induction, deduction

Population

Deduction,
Probability theory

Sample

Induction, Statistics

Table 2. Tg, swelling and elastic modulus of semi-IPNs cured according to thermal treatment

Run	Independent Variable			Dependent Variable					
(IPN)	NCO/OH	PU reaction time	MMA	Tg (°C) [a]		Swelling (%)		E (MPa)[b]	
	(molar ratio)	(min)	(%)	TT1[c]	TT2[d]	TT1[c]	TT2[d]	TT1[c]	TT2[d]
1	1.2	30	20	- 25.0	- 23.8	22.6	23.1	1.48	0.94
2	1.2	30	60	- 21.7	- 24.6	17.4	17.0	6.49	1.41
3	1.2	120	20	- 20.6	- 26.8	22.2	21.3	1.62	1.32
4	1.2	120	60	- 17.7	- 20.4	13.3	18.1	8.37	3.52
5	1.6	30	20	- 9.2	- 8.5	19.8	19.1	3.68	3.16
6	1.6	30	60	- 7.7	- 9.0	15.5	17.4	7.12	5.74
7	1.6	120	20	- 6.8	- 6.5	20.2	15.6	3.62	3.12
8	1.6	120	60	- 6.9	- 8.3	15.1	15.5	7.19	3.63
9	1.4	75	40	- 9.7	- 13.5	14.5	18.3	2.51	2.24
10	1.4	75	40	- 9.8	- 13.4	16.4	18.9	2.83	2.79
11	1.4	75	40	- 12.7	- 18.5	18.3	18.8	3.89	2.06
12	1.4	75	40	- 10.6	- 9.2	18.8	18.5	3.06	2.85

a) Tg obtained by DSC;
b) Modulus elastic obtained by DMA according ASTM–D882–91;
c) Thermal treatment TT1, oven 24 h at 70 °C and 4 h at 120 °C;
d) Thermal treatment TT2, room temperature 28 h.

What is data analysis?

data manipulation arbitrary origin and scale problems of space and time; data filtering, cleaning, outlier detection; reshaping data; import/export to common formats

data plotting bar charts, histograms, scatter plots, time series plots, maps, ...

data summarising descriptive statistics (mean, standard deviation, range, ...)

inference inferring population characteristics based on limited sample data (interval estimation, hypothesis testing, modelling)

Statistics provides generic theory, language and tools to perform data analysis.

What is data analysis?

data manipulation arbitrary origin and scale problems of space and time; data filtering, cleaning, outlier detection; reshaping data; import/export to common formats

data plotting bar charts, histograms, scatter plots, time series plots, maps, ...

data summarising descriptive statistics (mean, standard deviation, range, ...)

inference inferring population characteristics based on limited sample data (interval estimation, hypothesis testing, modelling)

Statistics provides generic theory, language and tools to perform data analysis.

What is data analysis?

data manipulation arbitrary origin and scale problems of space and time; data filtering, cleaning, outlier detection; reshaping data; import/export to common formats

data plotting bar charts, histograms, scatter plots, time series plots, maps, ...

data summarising descriptive statistics (mean, standard deviation, range, ...)

inference inferring population characteristics based on limited sample data (interval estimation, hypothesis testing, modelling)

Statistics provides generic theory, language and tools to perform data analysis.

What is data analysis?

data manipulation arbitrary origin and scale problems of space and time; data filtering, cleaning, outlier detection; reshaping data; import/export to common formats

data plotting bar charts, histograms, scatter plots, time series plots, maps, ...

data summarising descriptive statistics (mean, standard deviation, range, ...)

inference inferring population characteristics based on limited sample data (interval estimation, hypothesis testing, modelling)

Statistics provides generic theory, language and tools to perform data analysis.

ifgi

What is data analysis?

data manipulation arbitrary origin and scale problems of space and time; data filtering, cleaning, outlier detection; reshaping data; import/export to common formats

data plotting bar charts, histograms, scatter plots, time series plots, maps, ...

data summarising descriptive statistics (mean, standard deviation, range, ...)

inference inferring population characteristics based on limited sample data (interval estimation, hypothesis testing, modelling)

Statistics provides generic theory, language and tools to perform data analysis.

What is data analysis?

data manipulation arbitrary origin and scale problems of space and time; data filtering, cleaning, outlier detection; reshaping data; import/export to common formats

data plotting bar charts, histograms, scatter plots, time series plots, maps, ...

data summarising descriptive statistics (mean, standard deviation, range, ...)

inference inferring population characteristics based on limited sample data (interval estimation, hypothesis testing, modelling)

Statistics provides generic theory, language and tools to perform data analysis.

What do we mean and what is generally meant by geostatistics?

Wide sense (uni-muenster): the statistical analysis of geoscientific data (summarising, plotting, inference, modelling)

Narrow sense the modelling of spatial or spatio-temporal data by using *random functions*, modelling such functions from sample data, estimation (spatial interpolation) and simulation of these functionals.

To start with narrow-sense geostatistics, we first need an introduction to (wide sense) geostatistics. My second year course *Modelling Spatial and Temporal Processes* will dive (among others) further into narrow-sense geostatistics.

What do we mean and what is generally meant by geostatistics?

Wide sense (uni-muenster): the statistical analysis of geoscientific data (summarising, plotting, inference, modelling)

Narrow sense the modelling of spatial or spatio-temporal data by using *random functions*, modelling such functions from sample data, estimation (spatial interpolation) and simulation of these functionals.

To start with narrow-sense geostatistics, we first need an introduction to (wide sense) geostatistics. My second year course *Modelling Spatial and Temporal Processes* will dive (among others) further into narrow-sense geostatistics.

ifgi

What do we mean and what is generally meant by geostatistics?

Wide sense (uni-muenster): the statistical analysis of geoscientific data (summarising, plotting, inference, modelling)

Narrow sense the modelling of spatial or spatio-temporal data by using *random functions*, modelling such functions from sample data, estimation (spatial interpolation) and simulation of these functionals.

To start with narrow-sense geostatistics, we first need an introduction to (wide sense) geostatistics. My second year course *Modelling Spatial and Temporal Processes* will dive (among others) further into narrow-sense geostatistics.

ifgi

What do we mean and what is generally meant by geostatistics?

Wide sense (uni-muenster): the statistical analysis of geoscientific data (summarising, plotting, inference, modelling)

Narrow sense the modelling of spatial or spatio-temporal data by using *random functions*, modelling such functions from sample data, estimation (spatial interpolation) and simulation of these functionals.

To start with narrow-sense geostatistics, we first need an introduction to (wide sense) geostatistics. My second year course *Modelling Spatial and Temporal Processes* will dive (among others) further into narrow-sense geostatistics.

Why we use

- R is open source, `http://www.r-project.org/`
- R implements a programming language (S)
- R is extendible
- research should be reproducible

Why we use R

- R is open source, `http://www.r-project.org/`
- R implements a programming language (S)
- R is extendible
- research should be reproducible

ifgi

Why we use

- R is open source, `http://www.r-project.org/`
- R implements a programming language (S)
- R is extendible
- research should be reproducible

ifgi

Why we use

- R is open source, `http://www.r-project.org/`
- R implements a programming language (S)
- R is extendible
- research should be reproducible

Why we use open source software (OSS)

- OSS is not necessarily better (often worse), but it can be *verified* and *improved*, and thus leads to more control.
- Commercial vendors may stop developing, or change versions in an incompatible way.
- Commercial statistical software (SPSS, SAS, S-Plus, ...) is usually cheap for universities, but often expensive (€10.000+) for companies
- User support is often much faster, public, and provided by developers
- Absence of a financial barriers invites the reproduction of research results
- A large user base leads to well-tested and well-trusted code (this is more true for OSS than for other software)
- With open source in mind, people code differently.

Why we use open source software (OSS)

- OSS is not necessarily better (often worse), but it can be *verified* and *improved*, and thus leads to more control.
- Commercial vendors may stop developing, or change versions in an incompatible way.
- Commercial statistical software (SPSS, SAS, S-Plus, ...) is usually cheap for universities, but often expensive (€10.000+) for companies
- User support is often much faster, public, and provided by developers
- Absence of a financial barriers invites the reproduction of research results
- A large user base leads to well-tested and well-trusted code (this is more true for OSS than for other software)
- With open source in mind, people code differently.

ifgi

Why we use open source software (OSS)

- OSS is not necessarily better (often worse), but it can be *verified* and *improved*, and thus leads to more control.
- Commercial vendors may stop developing, or change versions in an incompatible way.
- Commercial statistical software (SPSS, SAS, S-Plus, ...) is usually cheap for universities, but often expensive (€10.000+) for companies
- User support is often much faster, public, and provided by developers
- Absence of a financial barriers invites the reproduction of research results
- A large user base leads to well-tested and well-trusted code (this is more true for OSS than for other software)
- With open source in mind, people code differently.

Why we use open source software (OSS)

- OSS is not necessarily better (often worse), but it can be *verified* and *improved*, and thus leads to more control.
- Commercial vendors may stop developing, or change versions in an incompatible way.
- Commercial statistical software (SPSS, SAS, S-Plus, ...) is usually cheap for universities, but often expensive (€10.000+) for companies
- User support is often much faster, public, and provided by developers
- Absence of a financial barriers invites the reproduction of research results
- A large user base leads to well-tested and well-trusted code (this is more true for OSS than for other software)
- With open source in mind, people code differently.

Why we use open source software (OSS)

- OSS is not necessarily better (often worse), but it can be *verified* and *improved*, and thus leads to more control.
- Commercial vendors may stop developing, or change versions in an incompatible way.
- Commercial statistical software (SPSS, SAS, S-Plus, ...) is usually cheap for universities, but often expensive (€10.000+) for companies
- User support is often much faster, public, and provided by developers
- Absence of a financial barriers invites the reproduction of research results
- A large user base leads to well-tested and well-trusted code (this is more true for OSS than for other software)
- With open source in mind, people code differently.

ifgi

Why we use open source software (OSS)

- OSS is not necessarily better (often worse), but it can be *verified* and *improved*, and thus leads to more control.
- Commercial vendors may stop developing, or change versions in an incompatible way.
- Commercial statistical software (SPSS, SAS, S-Plus, ...) is usually cheap for universities, but often expensive (€10.000+) for companies
- User support is often much faster, public, and provided by developers
- Absence of a financial barriers invites the reproduction of research results
- A large user base leads to well-tested and well-trusted code (this is more true for OSS than for other software)
- With open source in mind, people code differently.

Why we use open source software (OSS)

- OSS is not necessarily better (often worse), but it can be *verified* and *improved*, and thus leads to more control.
- Commercial vendors may stop developing, or change versions in an incompatible way.
- Commercial statistical software (SPSS, SAS, S-Plus, ...) is usually cheap for universities, but often expensive (€10.000+) for companies
- User support is often much faster, public, and provided by developers
- Absence of a financial barriers invites the reproduction of research results
- A large user base leads to well-tested and well-trusted code (this is more true for OSS than for other software)
- With open source in mind, people code differently.

R implements a programming language called S

- Developed at AT&T Bell Labs (C++), 1980s, by Becker, Wilks, Chambers,
- S is a language for data analysis, a special-purpose language rather than a general purpose language.
- vector-oriented (just like e.g. matlab)
- Re-developed in 2000 (S4)
- Object-orientation, methods-based rather than class-based (quite different from e.g. Java, Python)
- efficient to reach data analysis goals, using tested and trusted code.
- scaling up of analysis: commands → script → function → new command; UseR!s often become programmers.

R implements a programming language called S

- Developed at AT&T Bell Labs (C++), 1980s, by Becker, Wilks, Chambers,
- S is a language for data analysis, a special-purpose language rather than a general purpose language.
- vector-oriented (just like e.g. matlab)
- Re-developed in 2000 (S4)
- Object-orientation, methods-based rather than class-based (quite different from e.g. Java, Python)
- efficient to reach data analysis goals, using tested and trusted code.
- scaling up of analysis: commands → script → function → new command; UseR!s often become programmers.

R implements a programming language called S

- Developed at AT&T Bell Labs (C++), 1980s, by Becker, Wilks, Chambers,
- S is a language for data analysis, a special-purpose language rather than a general purpose language.
- vector-oriented (just like e.g. matlab)
- Re-developed in 2000 (S4)
- Object-orientation, methods-based rather than class-based (quite different from e.g. Java, Python)
- efficient to reach data analysis goals, using tested and trusted code.
- scaling up of analysis: commands → script → function → new command; UseR!s often become programmers.

R implements a programming language called S

- Developed at AT&T Bell Labs (C++), 1980s, by Becker, Wilks, Chambers,
- S is a language for data analysis, a special-purpose language rather than a general purpose language.
- vector-oriented (just like e.g. matlab)
- Re-developed in 2000 (S4)
- Object-orientation, methods-based rather than class-based (quite different from e.g. Java, Python)
- efficient to reach data analysis goals, using tested and trusted code.
- scaling up of analysis: commands → script → function → new command; UseR!s often become programmers.

R implements a programming language called S

- Developed at AT&T Bell Labs (C++), 1980s, by Becker, Wilks, Chambers,
- S is a language for data analysis, a special-purpose language rather than a general purpose language.
- vector-oriented (just like e.g. matlab)
- Re-developed in 2000 (S4)
- Object-orientation, methods-based rather than class-based (quite different from e.g. Java, Python)
- efficient to reach data analysis goals, using tested and trusted code.
- scaling up of analysis: commands → script → function → new command; UseR!s often become programmers.

R implements a programming language called S

- Developed at AT&T Bell Labs (C++), 1980s, by Becker, Wilks, Chambers,
- S is a language for data analysis, a special-purpose language rather than a general purpose language.
- vector-oriented (just like e.g. matlab)
- Re-developed in 2000 (S4)
- Object-orientation, methods-based rather than class-based (quite different from e.g. Java, Python)
- efficient to reach data analysis goals, using tested and trusted code.
- scaling up of analysis: commands → script → function → new command; UseR!s often become programmers.

R implements a programming language called S

- Developed at AT&T Bell Labs (C++), 1980s, by Becker, Wilks, Chambers,
- S is a language for data analysis, a special-purpose language rather than a general purpose language.
- vector-oriented (just like e.g. matlab)
- Re-developed in 2000 (S4)
- Object-orientation, methods-based rather than class-based (quite different from e.g. Java, Python)
- efficient to reach data analysis goals, using tested and trusted code.
- scaling up of analysis: commands → script → function → new command; UseR!s often become programmers.

Matrix, index, replacement

```
> a = matrix(1:4, 2, 2, byrow = TRUE)
> a

     [,1] [,2]
[1,]    1    2
[2,]    3    4

> a[1, 2]

[1] 2

> a[1, 2] = 10
> a

     [,1] [,2]
[1,]    1   10
[2,]    3    4
```

Two-group data, two shapes

```
> A = c(1, 3, 2, 5)
> B = c(9, 12, 14)
> list(A = A, B = B)

$A
[1] 1 3 2 5

$B
[1]  9 12 14

> data.frame(value = c(A, B), group = c(rep("A", length(A)),
+     rep("B", length(B))))

  value group
1     1     A
2     3     A
3     2     A
4     5     A
5     9     B
6    12     B
7    14     B
```

R is extendible

- more than 2300 (Apr 2010) user-contributed packages on CRAN `http://cran.r-project.org/` (used by R-core to check R development)
- packages contain tests, allow for unit-testing
- well-documented foreign-language API
- R glues well to other environments (COM, ogr/gdal, GRASS GIS, data bases, ...)

R is extendible

- more than 2300 (Apr 2010) user-contributed packages on CRAN `http://cran.r-project.org/` (used by R-core to check R development)
- packages contain tests, allow for unit-testing
- well-documented foreign-language API
- R glues well to other environments (COM, ogr/gdal, GRASS GIS, data bases, ...)

R is extendible

- more than 2300 (Apr 2010) user-contributed packages on CRAN `http://cran.r-project.org/` (used by R-core to check R development)
- packages contain tests, allow for unit-testing
- well-documented foreign-language API
- R glues well to other environments (COM, ogr/gdal, GRASS GIS, data bases, ...)

R is extendible

- more than 2300 (Apr 2010) user-contributed packages on CRAN `http://cran.r-project.org/` (used by R-core to check R development)
- packages contain tests, allow for unit-testing
- well-documented foreign-language API
- R glues well to other environments (COM, ogr/gdal, GRASS GIS, data bases, ...)

R is full of research challenges

- developing and distributing new analysis algorithms
- dealing with massive data sets
- service-oriented architectures
- multi-threading, parallel/GRID computation
- analysis of spatial and spatio-temporal data

R is full of research challenges

- developing and distributing new analysis algorithms
- dealing with massive data sets
- service-oriented architectures
- multi-threading, parallel/GRID computation
- analysis of spatial and spatio-temporal data

R is full of research challenges

- developing and distributing new analysis algorithms
- dealing with massive data sets
- service-oriented architectures
- multi-threading, parallel/GRID computation
- analysis of spatial and spatio-temporal data

R is full of research challenges

- developing and distributing new analysis algorithms
- dealing with massive data sets
- service-oriented architectures
- multi-threading, parallel/GRID computation
- analysis of spatial and spatio-temporal data

R is full of research challenges

- developing and distributing new analysis algorithms
- dealing with massive data sets
- service-oriented architectures
- multi-threading, parallel/GRID computation
- analysis of spatial and spatio-temporal data

Reproducible research

- Good policy is built upon state-of-the art scientific knowledge
- scientific knowledge progresses by peer-review
- peer-review should not only address the research results (papers) but also the reproduction of experiments and analysis of data.
- example: clinical trials, particle physics, climate change
- When analysis steps are written in a (clean) script file, understanding, verifying and reproducing the analysis is easier then when these steps are described as a large sequence of mouse clicks and moves.

Reproducible research

- Good policy is built upon state-of-the art scientific knowledge
- scientific knowledge progresses by peer-review
- peer-review should not only address the research results (papers) but also the reproduction of experiments and analysis of data.
- example: clinical trials, particle physics, climate change
- When analysis steps are written in a (clean) script file, understanding, verifying and reproducing the analysis is easier then when these steps are described as a large sequence of mouse clicks and moves.

Reproducible research

- Good policy is built upon state-of-the art scientific knowledge
- scientific knowledge progresses by peer-review
- peer-review should not only address the research results (papers) but also the reproduction of experiments and analysis of data.
- example: clinical trials, particle physics, climate change
- When analysis steps are written in a (clean) script file, understanding, verifying and reproducing the analysis is easier then when these steps are described as a large sequence of mouse clicks and moves.

ifgi

Reproducible research

- Good policy is built upon state-of-the art scientific knowledge
- scientific knowledge progresses by peer-review
- peer-review should not only address the research results (papers) but also the reproduction of experiments and analysis of data.
- example: clinical trials, particle physics, climate change
- When analysis steps are written in a (clean) script file, understanding, verifying and reproducing the analysis is easier then when these steps are described as a large sequence of mouse clicks and moves.

Reproducible research

- Good policy is built upon state-of-the art scientific knowledge
- scientific knowledge progresses by peer-review
- peer-review should not only address the research results (papers) but also the reproduction of experiments and analysis of data.
- example: clinical trials, particle physics, climate change
- When analysis steps are written in a (clean) script file, understanding, verifying and reproducing the analysis is easier then when these steps are described as a large sequence of mouse clicks and moves.