How to Write Scientific Papers

UNI FREIBURG

University of Freiburg, Germany

Stefan Feuerriegel

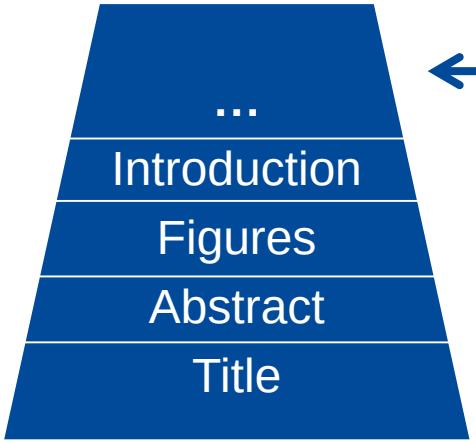
Chair for Information Systems Research
April 2015

http://www.nature.com/scitable/content/ne0000/ne0000/ne0000/ne0000/14239512/ECS_scientific-papers_ksm.jpg

But, then, how do you write papers?

The Reader's Perspective

- Researchers interest different levels of your paper



What nobody reads

What most read

Your task is to get your message across

Agenda

1. Paper Writing Process
2. Searching & Structuring Literature
3. Organizing the Paper
4. Writing the Paper

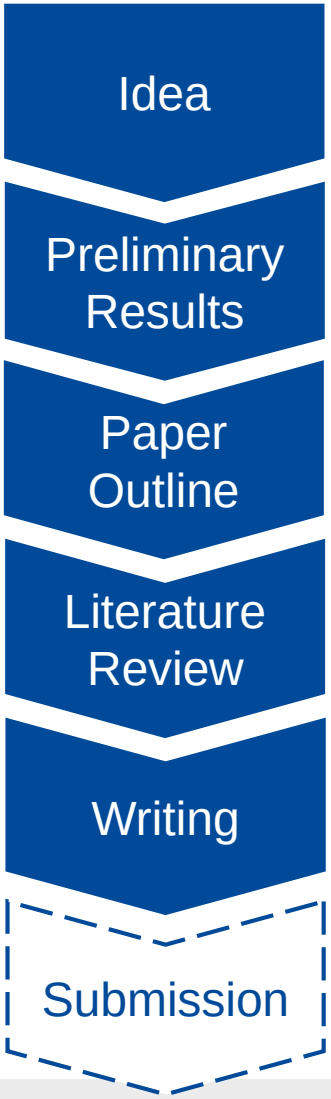
This presentation is based on author's experience in informatics-related studies. There is no single "right" way

Agenda

UNI FREIBURG

1. Paper Writing Process
2. Searching & Structuring Literature
3. Organizing the Paper
4. Writing the Paper

Writing Papers: The Process



Formulate research question and check novelty

Verify research questions early; continue if sound (≤ 3 months)

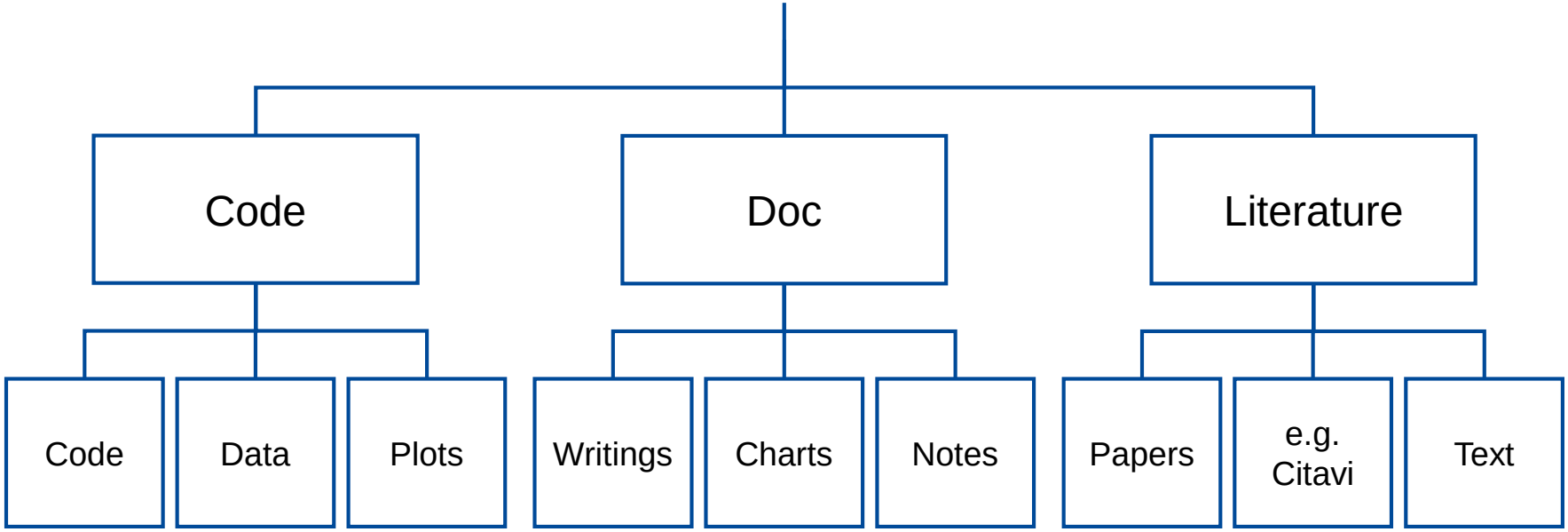
Specify sections and content, possibly at paragraph level (1/2 day)

Reading & structuring papers, possibly in table format (1 week)

Iterative revisions, plotting, and getting final results (1–6 weeks)

Storing your Data

- Use of repository is highly recommended, e.g. SVN, git
- Reasons: backup, roll backs, collaboration
- Contains **everything** except temporary files (exe, Rdata, …)
- Store also intermediate results (≈ 1 check-in per hour)
- Folders: branch (side-line), trunk (main), tags (marked versions)



Agenda

1. Paper Writing Process
2. Searching & Structuring Literature
3. Organizing the Paper
4. Writing the Paper

Searching Literature

Reasons for Referencing

- Similar research
- Proof of relevance
- Proof of novelty
  (by pointing out differences)
- Same methodology
- Links for background search
- Theories for discussion

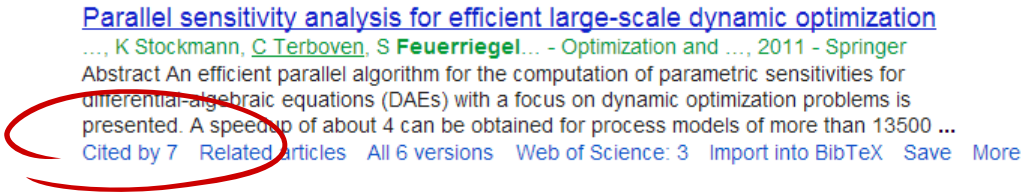
Sources

- **Google Scholar**
- Web of Knowledge
- ScienceDirect
- Journals:
  Elsevier, Springer, IEEE, ACM
- **Subscribe** to Google Scholar alert for relevant keywords

Traversing the Citation Graph

Try to find all relevant papers:
→ missing one is likely to lead to a reject

1. Read the 3–7 closest papers (P) carefully
   Mark: methodological details, text worth quoting, etc.
   Check cross references in (P), especially in related work section
2. Use Google Scholar to search papers citing (P)



3. Search other papers from the authors of (P)

Collecting Literature

Organizing

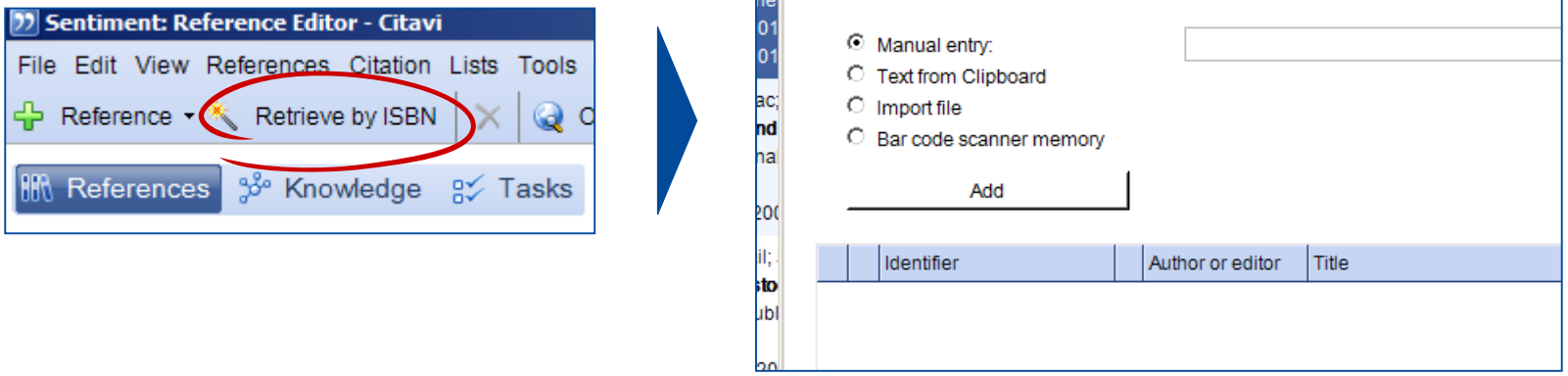
- 10–40 references in each paper, but scanning up to 100 → good organization is essential
- Collect PDFs on local computer for faster access
- Collect also quotations and notes if helpful
- Use a software for literature organization

Tools

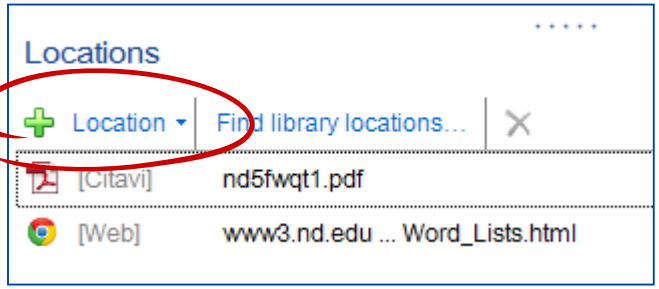
- **Citavi** is powerful, but Windows only (recommended)
- EndNote, Zotero are common, but less powerful
- MS Word possibly, but cannot organize references
- Mendeley is web-based, but citation styles troublesome
- BibDesk can be an alternative on Mac OS

Citavi: Importing

- Fastest import: Copy DOI and let Citavi retrieve data
- Always double check (e.g. ACM fails)

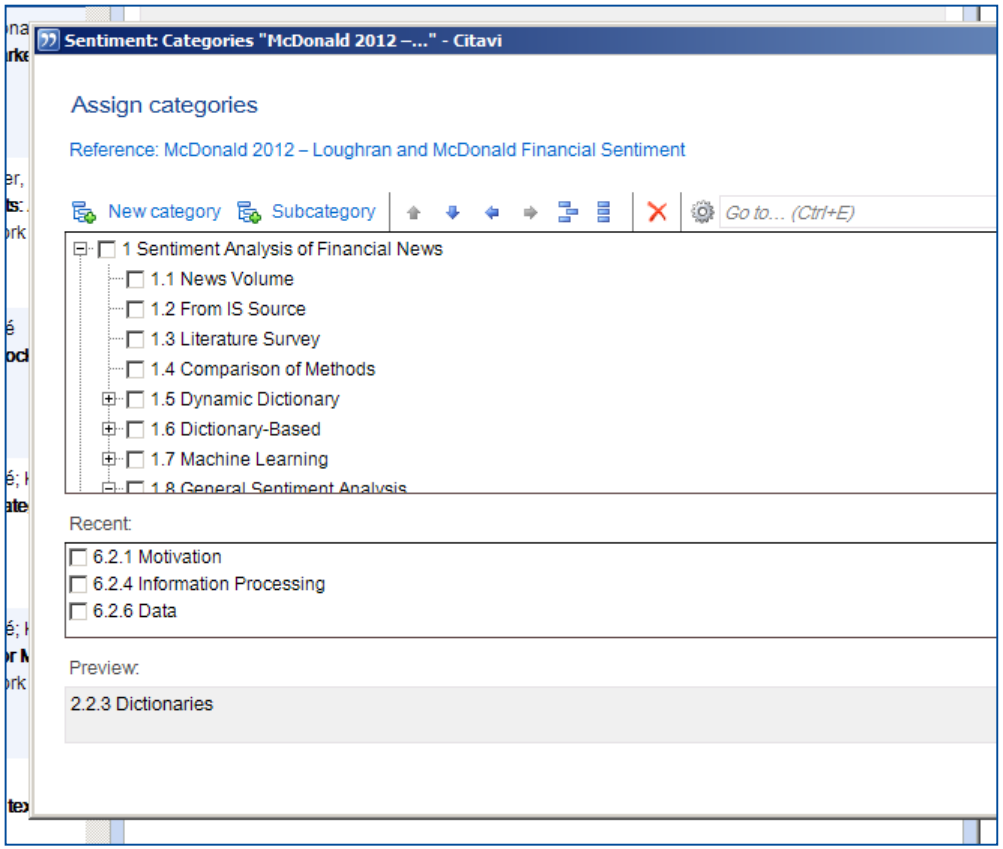


- Copy PDFs into Citavi



Citavi: Categorizing

- Use categories for structuring



“Good” References

Try citing references from “good” sources
→ proxy for scientific validity and impact

“good”

English

“avoid”

German, etc.

books (→ background)

web (data sources only)

Asian or BSc authors

conferences

journals

highly ranking

same domain

not ranked

Journal Rankings	▪ Impact Factor ▪ Handelsblatt: http://bit.ly/bwljournals2012 ▪ VHB: http://vhbonline.org/service/jourqual/

Agenda

UNI FREIBURG

1 Paper Writing Process

2 Searching & Structuring Literature

3 Structuring the Paper

4 Writing the Paper

Outline

Section	Content
Abstract	What is the content in a nutshell?
Introduction	What is the problem?
Related Work	What are differences/similarities?
Methods	How is the problem solved?
Results	What are the findings?
Discussion	What does it mean?
Conclusion & Outlook	
Acknowledgements	(journals only)
Bibliography	What are the works referred to?
Appendices (optional)	Extra information

Materials
(or own section)

Achieving a Line of Thought

Without a Message

Introduction

- Regression models cannot show causality between news sentiment and stock market reaction

Problem Statement

- Use IV regression instead

Related Work

- Related price models

With a Line of Thought

Introduction

- News have impact on stock markets

Problem Statement

- Causality is frequently assumed, but rarely investigated (Loughran, 2011)

Research Question

- Show causality using IV regression

Related Work

- Related price on linking news sentiment and stock prices
- Reference price models

Abstract

- Short summary that must stand on its own
- Important: basis for decision upon reading the paper
- Length: ≤ 150 words in 1 paragraph
- Contains no references

Structure

1. Motivate relevance of problem
   (except highly-focused outlets in natural sciences)
2. State the problem briefly
3. Name methodology
4. Summarize outcome, give quantitative results

Abstract

Example 1

Check abstract here:

http://arxiv.org/abs/1403.6426

→ From Peise et al. (2014): High Performance Solutions for Big-data GWAS, Parallel Computing.

Abstract

Example 2

Since the liberalization of European electricity markets, stakeholders can actively participate in the trading of electricity. Successful participation in such markets requires an accurate forecast of future electricity prices. However, as large volumes of energy from renewable sources are fed into the system, electricity prices are highly volatile. While recent approaches put a strong focus on models from time series analysis using only historic prices, they neglect the influence of exogenous predictors. This paper improves state of the art research by identifying and including a set of exogenous predictors, namely, both expected solar and expected wind power generation to model the supply side as well as expected electricity load to model the demand side. Consequently, we evaluate our forecasting models by using a two-pronged approach. First, we show that these externals decrease root mean squared errors by between 3.37% and 9.86%. Second, we apply a Diebold-Mariano test to prove statistically that the forecasting accuracy of the models including exogenous predictors is superior.

Introduction

- Most read introduction only → important to get message across
- Helpful tools for structuring
  - Using **quotations** from top-tier papers to prove relevance/novelty
  - Visualizing relationships through **graphics**
  - **Research questions** or hypothesis (as own highlighted paragraph)
- Don't write "… is important", instead give arguments

Structure:
4 paragraphs

1. Motivate relevance of problem
   (probably give context)
2. State the problem briefly
3. Repeat contribution, e.g. as research questions
   (not in all journals, but recommended)
4. Outline (structure of section as story line)

Research Questions

Helpful when analyzing effects, not when proposing or comparing algorithms → use 3–4 of them

Examples

- Research Question 1: (a) To what extent are abnormal returns of oil driven by news sentiment? (b) To what extent are abnormal returns of gold driven by news sentiment?
- Research Question 2: To what extent are abnormal returns in commodity markets driven by news volume?
- Research Question 3: What is the influence of positive and negative news sentiment on commodity prices?

- H1: The stock market increases when social mood levels raises.
- H2: The stock market decreases in case of worsening financial crisis sentiment.

Outline

UNI FREIBURG

Example

The remainder of this paper is structured as follows. Section 2 provides a literature overview of publications forecasting electricity prices, where the majority of models ignores external impacts. To close this research gap, Section 3 utilizes both time series analysis and statistical learning to present models that incorporate exogenous predictors. These models are evaluated in Section 4, which, finally, reveals that external inputs improve forecasting accuracy at a statistically significant level.

- Not just name “Section 2 analyzes …”, but link sections
  → this explains and motivates the need for each section

Related Work

Goal

Major Make clear that research is relevant and novel

Minor State similar research to outline context

- Not just enumerate references, but summarize findings and differences to own study
- Structure also as table as helpful, visual tool to prove novelty
- Final paragraph is a wrap-up stressing that given the above references your research is novel and relevant
- Avoid definite statements, e.g. “this is new”
  → instead “we are not aware of similar research in the literature”
  or “to our best knowledge”

Tabular Presentation of Related Work

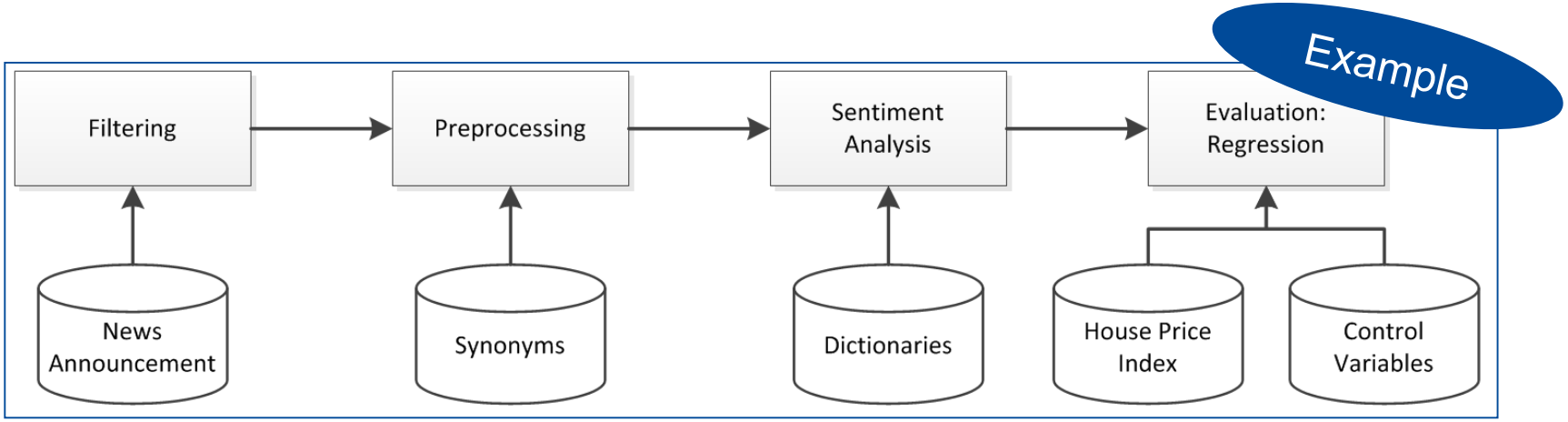
→ easy to identify what is different from previous research

Example

Reference	Market	Price	Model	Forecast Cnt	Rolling	Out-of-S	Load	Wind	Sol
Bello and Reneses (2013)	Spain	Daily average day-ahead	VECM	✗	✗	✓	✓	✓	✓
Contreras et al. (2003)	California, Spain	Hourly day-ahead	ARIMA, ARIMAX	✓	✓	✓	✓	✗	✗
Cruz et al. (2011)	Spain	Hourly day-ahead	ARIMA, dynamic regression, ANN	✓	✗	✓	✓	✓	✗
Erni (2012)	EPEX	Hourly day-ahead	GARCH, threshold regression, time-varying parameter regression	✗	✗	✓	✓	✓	✗
Huurman et al. (2012)	Nord-Pool	Daily average day-ahead	ARIMA, ARIMAX, ARIMAX-GARCH	✓	✗	✓	✗	✗	✗
Keles et al. (2013)	EEX	Hourly day-ahead	Own	✓	✗	✗	✗	✓	✗
Knittel and Roberts (2005)	Northern California	Hourly	Mean-reverting process, jump diffusion, EGARCH, ARMAX	✗	✗	✓	✓	✗	✗
Kosater (2006)	EEX	Hourly spot	Markov-regime switching	✗	✗	✓	✗	✗	✗
Kristiansen (2012)	Nord-Pool	Hourly day-ahead	ARMAX	✗	✗	✓	✓	✓	✗
Misiorek et al. (2006)	California	Hourly day-ahead	ARX, ARX-GARCH, threshold autoregression, Markov-regime switching	✓	✗	✓	✓	✗	✗
Nogales et al. (2002)	Spain, California	Hourly day-ahead	Dynamic regression, transfer function	✗	✗	✓	✓	✗	✗
Nogales and Conejo (2006)	PJM	Hourly day-ahead	Dynamic regression, transfer function	✓	✗	✓	✓	✗	✗
Szkuta et al. (1999)	Victoria	Half-hourly day-ahead	ANN	✗	✗	✓	✓	✗	✗
Weron and Misiorek (2008)	California, Nord-Pool	Hourly day-ahead	AR, ARX, spike model, regime-switching, mean-reverting jump diffusion	✓	✓	✓	✓	✗	✗
Xu and Niimura (2004)	PJM	Hourly day-ahead	Wavelet transformation	✗	✗	✓	✓	✗	✗
This paper	EPEX	Hourly day-ahead	ARMA, ARMAX, ANN	✓	✓	✓	✗	✓	✓

Methodology

- Pay attention to formulae
  → even little errors seem unprofessional
- Visualize via flow diagram if helpful



Results

- When highlighting findings, use 3–4 matching your research questions
- Find appropriate visualization (tables, charts)

Ideas for Discussion

- Limitations (→ also in Outlook)
- Managerial Implications
- Future Impact

Conclusion & Outlook

Structure:
3 paragraphs

1. Repeat problem and its relevance
2. Repeat contribution (plus quantitative results)
3. Give outlook on further research steps

Agenda

UNI FREIBURG

1 Paper Writing Process

2 Searching & Structuring Literature

3 Structuring the Paper

4 Writing the Paper

Figures

- Plots must be printer-friendly
  → e.g. black/white coloring or solid/dashed line
- Name axes with units, as well as legend
- Always state the message of what you see in the text
- Recommended also to explain plot structure in the text
  → including which data rows, possible trend or outliers
- Reference figure always in the text (“… in Figure 3”)

Detailed Captions

- Aim: Most scan papers, thus, plot must be understandable without surrounding text
- Name axes and data rows
- Give details on set-up of e.g. experiments

Captions

Examples

- Related literature on forecasting electricity prices using exogenous predictors
- Comparison of forecasting accuracy in terms of root mean squared error (RMSE), measured from Feb 01, 2010 to March 21, 2012 across various models
- Scatterplot of hourly day-ahead electricity prices and expected generation of wind (left) and solar (right) power with LOWESS trend line from … to …
- Pooled regression comparing news reception in bullish and bearish markets
- Demand Response potential through load shifting in Germany (Klobasa, 2007); scaled according to retailer's electricity demand

Writing Process

- Structure is important → late changes are expensive

Prepare a table with the structure and use this also to note the content of each section with a few bullet points

Make bullet points for each paragraph with a message
→ 1 paragraph = 1 message, and consistent flow within paragraph

Write quickly, to get a draft

Then improve the writing iteratively
→ follow advice in Gopen: The Science of Scientific Writing (1990)

Quotations

- Use direct quotations rarely, only around 0–5 per paper

Don't Use	▪ Mathematical definitions ▪ Short expressions
Only Use	▪ Trigger credibility of relevance in introduction ▪ Provide statement that the problem is not solved in literature ▪ In comparison with other papers/theories in discussion → strong statements from "big guys"

Beginning a Section

- Always separate two headlines by introductory sentences
- State briefly what you are doing at beginning of each section
  → Control and prepare what your reader expects next

Examples

- In this section, we present related literature grouped into two categories.
- To fully capture the concepts of …, we start by presenting different angles of the existing … approaches.
- This section introduces the theoretical background on modeling …
- In this section, we evaluate our proposed models in terms of ….

Simple Levers for Plain Language

Avoid long noun sequences
→ Estimating values of the coefficients of our econometric model is via OLS

Rewrite long insertions before a short main clause (rule of thumb: the longest part should follow the main verb)
→ In order to estimate the coefficients in our econometric model, OLS is used

Replace too frequent passive verbs with active voice
→ OLS performs the estimation of the coefficients in our econometric model

Benefit from the use of "we" to highlight your own work
→ We use OLS in order to estimate the coefficients in our econometric model

Use sentence connectors from a large variety for coherent text
→ e.g. furthermore, however, though, while, consequently, … (see lists online)

Since most of your readers are no native speakers, make your writings understandable

Hints

- Use understandable/clean English as native speakers are rare
- Ask native speaker for proof-reading
- Use highlighting where necessary (italics/emph only)
- Use high quality graphics, not pixelated
- Numbers: use 2–4 significant digits (except amount of money)

Dictionaries

- http://dict.leo.org for Germans to find alternatives
- http://ldoceonline.com to check the **meaning** and proper usage
- http://dictionary.com is a good alternative
- http://thesaurus.com to find **synonyms**