Application and available datasets for the analysis of distributional SD

A. Irpino, R. Verde

ESTP Cologne 14-16 May 2024

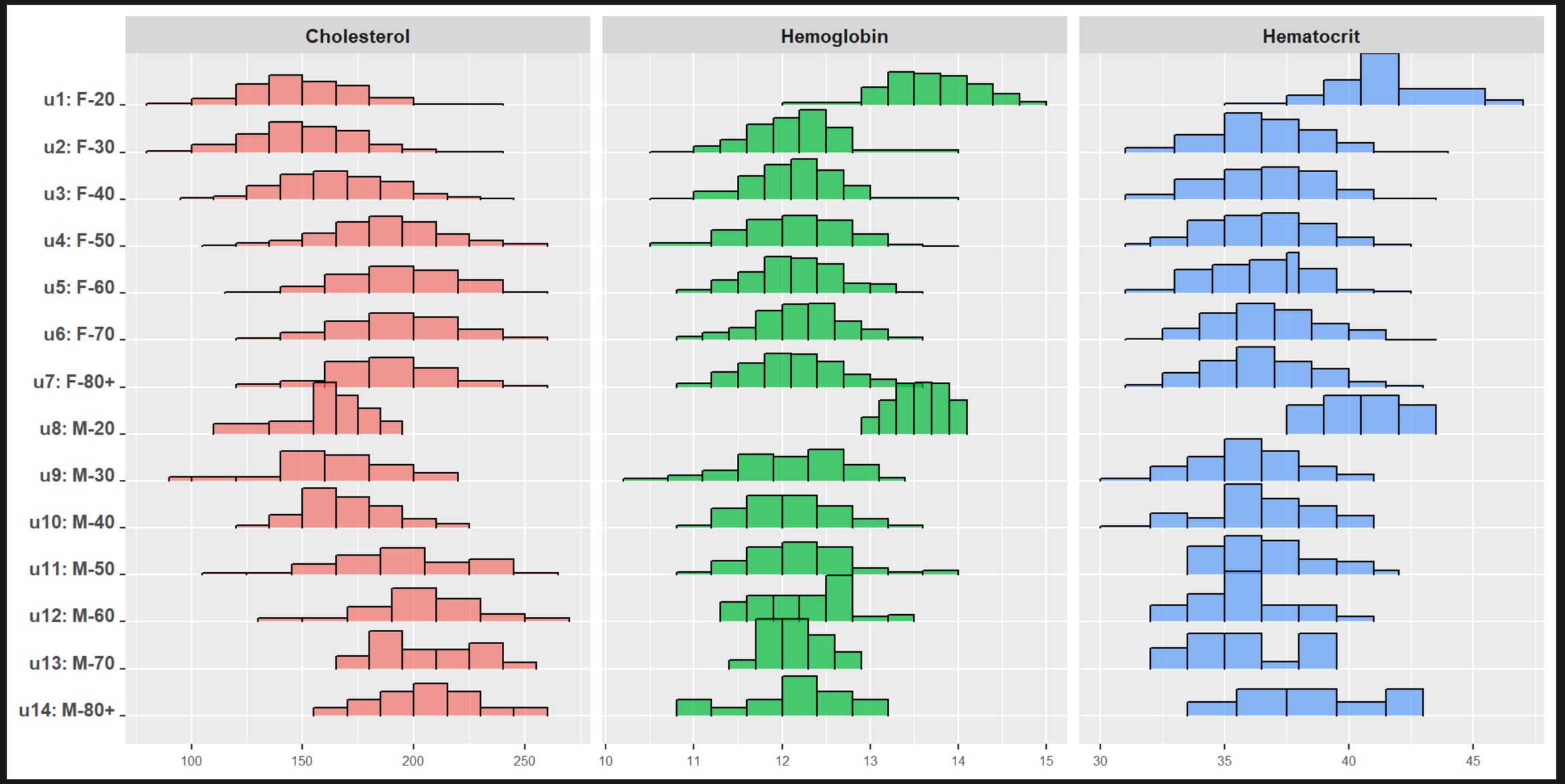
Datasets available for exercitating

Datasets contained in the `HistDAWass` package

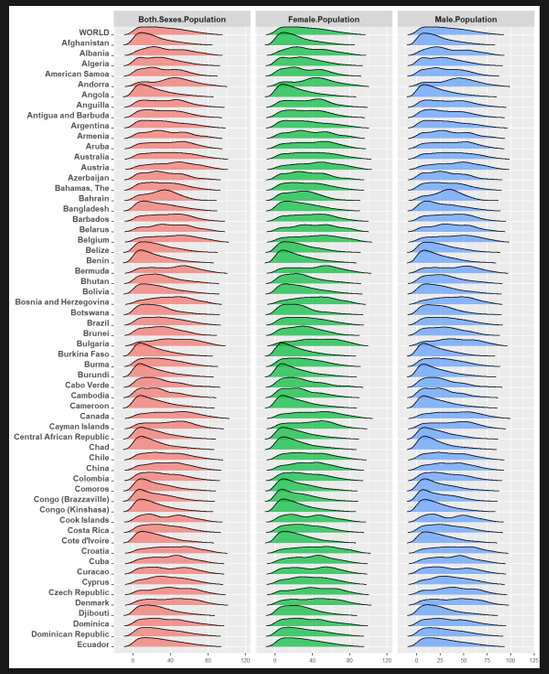
The `HistDAWass` package in R contains some distributional datasets generated from some books and from raw data from different contexts of application. Let's see some of them:

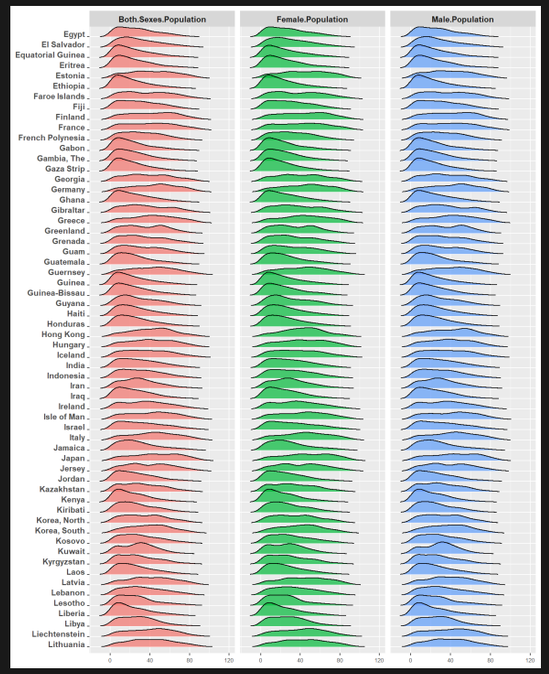
- `BLOOD` dataset, is the description of 14 typologies of patients in a hospital, the variables are histograms of the levels o Hematocrit, Hemoglobin and Cholesterol
- `Age_Pyramids_2014` Age pyramids of all the countries of the World in 2014
- `China_Month` Distributions of climatic variables for each month of 60 stations
- `China_Seas` Distributions of climatic variables for each season of 60 stations China
- `OzoneFull` Full Ozone dataset for Histogram data analysis

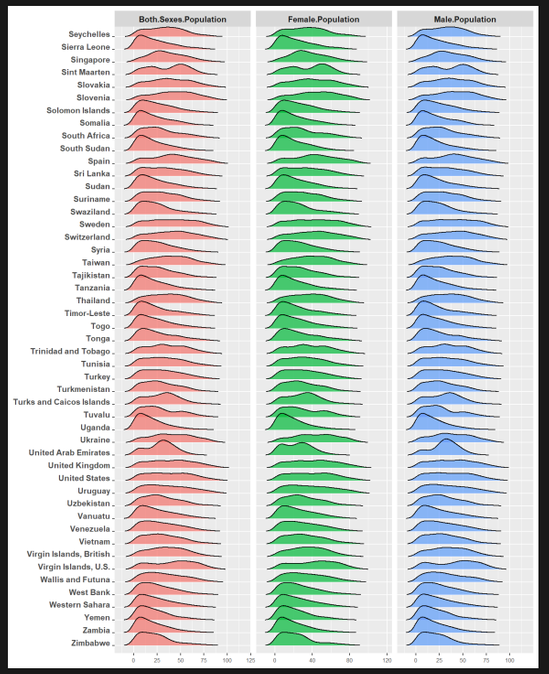
The BLOOD dataset



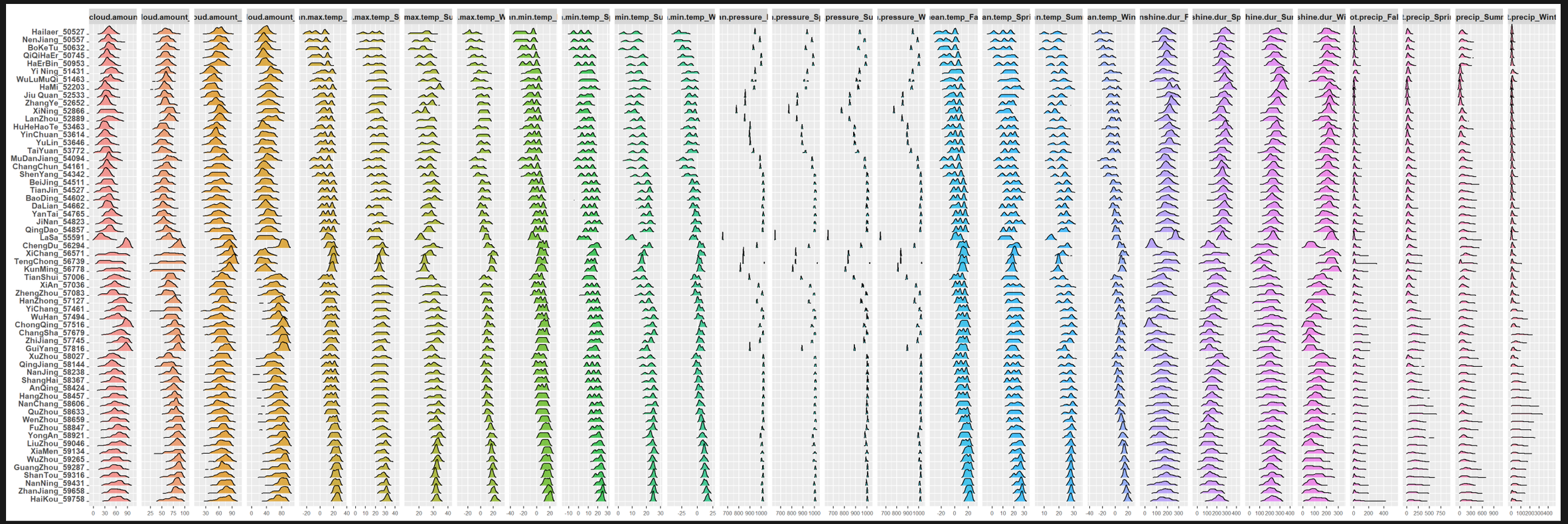
The AGE PYRAMIDS dataset (229 countries)

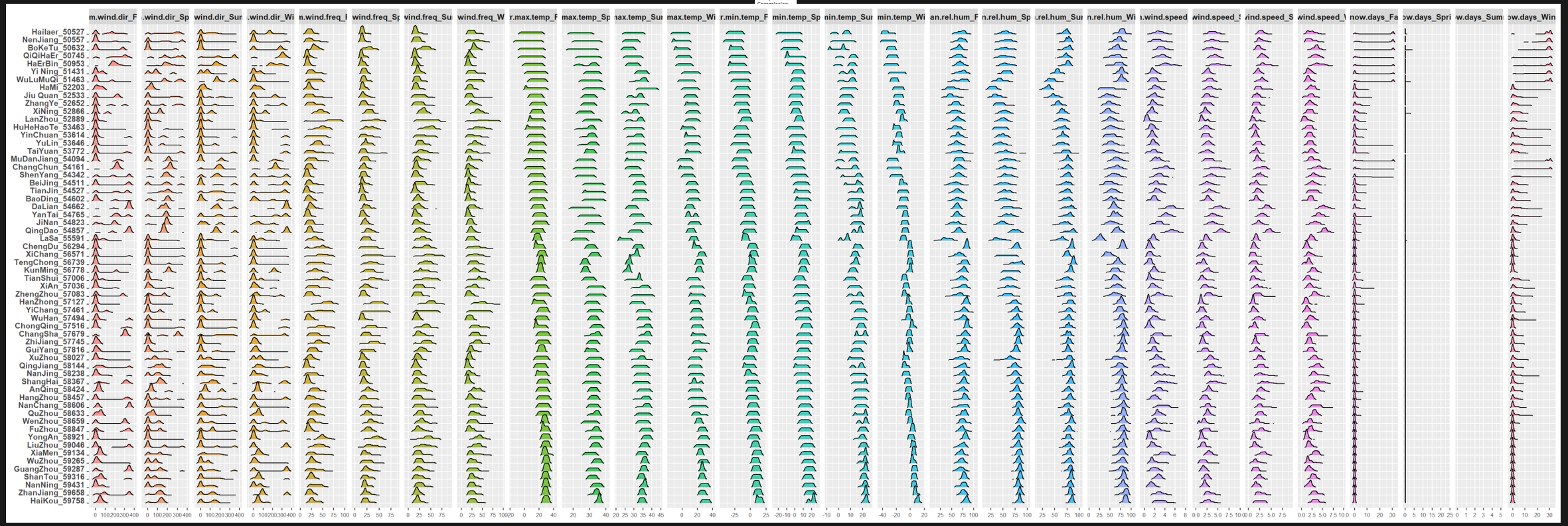




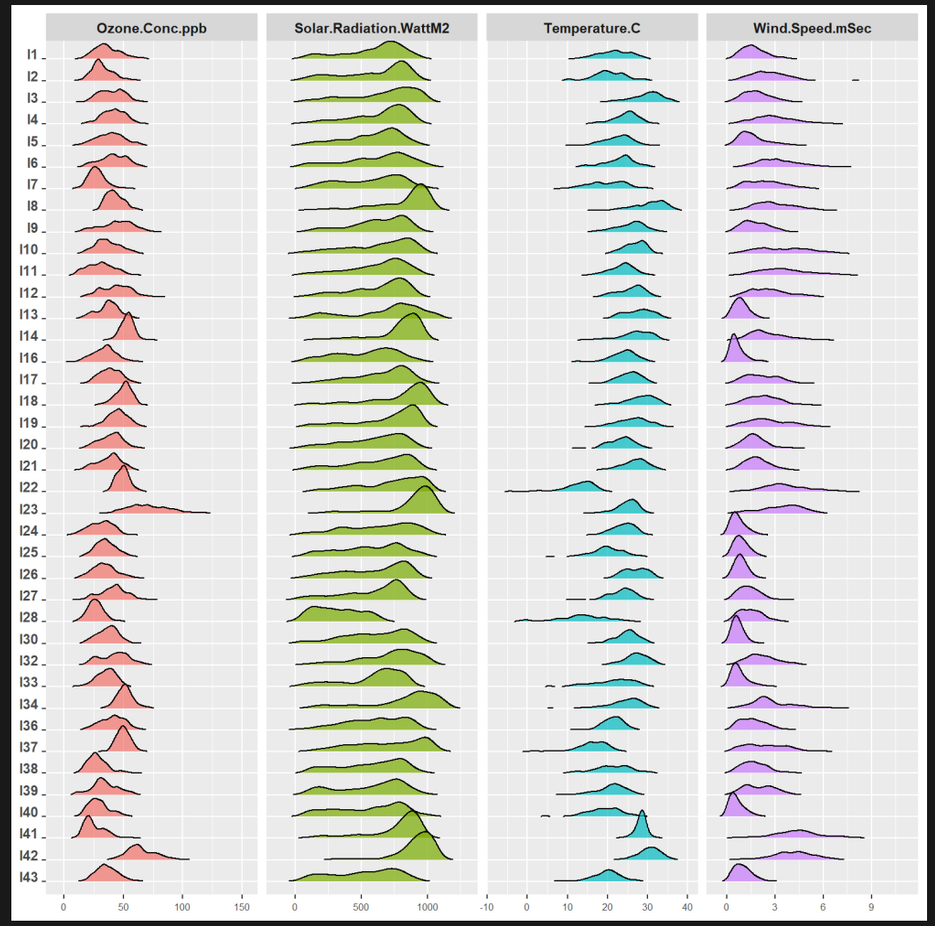


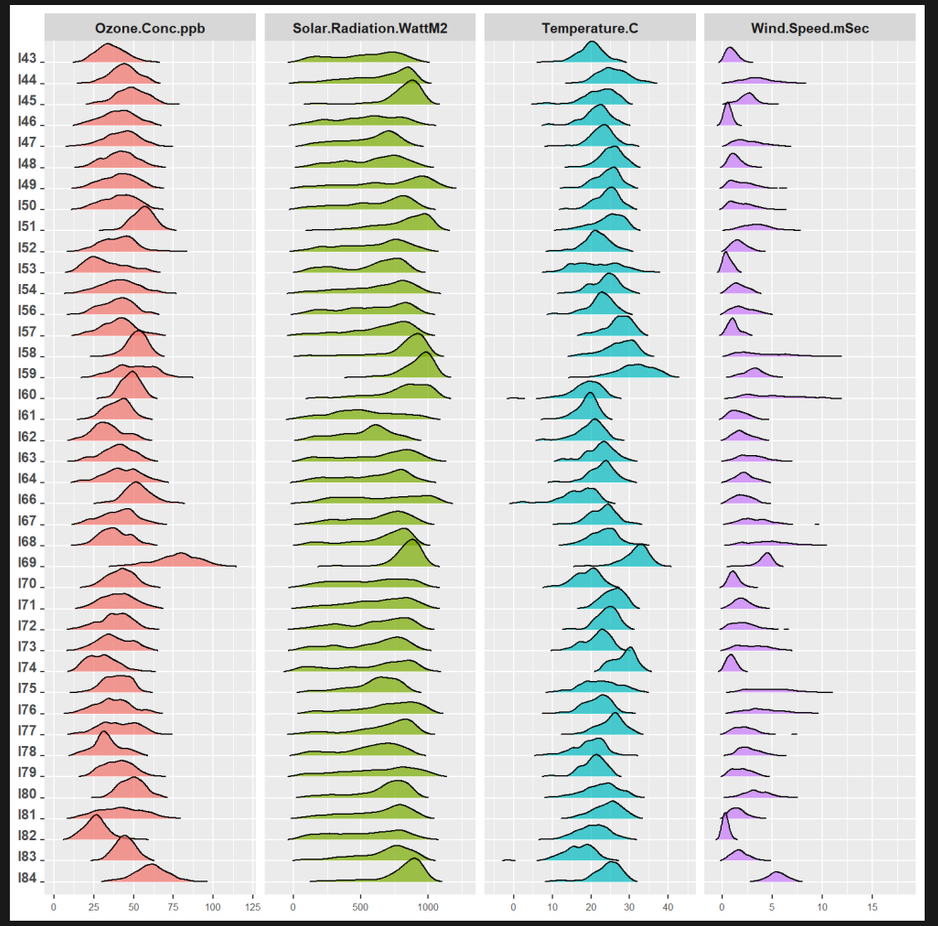
The CHINA season dataset 60 obs, 56 vars




Hailaer_50527
NenJiang_50557
BoKeTu_50632
QiQiHaEr_50745
HaErBin_50953
Yi Ning_51431
WuLuMuQi_51463
HaMi_52203
Jiu Quan_52533
ZhangYe_52652
XiNing_52866
LanZhou_52889
HuHeHaoTe_53463
YinChuan_53614
YuLin_53646
TaiYuan_53772
MuDanJiang_54094
ChangChun_54161
ShenYang_54342
BeiJing_54511
TianJin_54527
BaoDing_54602
DaLian_54662
YanTai_54765
JiNan_54823
QingDao_54857
LaSa_55591
ChengDu_56294
XiChang_56571
TengChong_56739
KunMing_56778
TianShui_57006
XiAn_57036
ZhengZhou_57083
HanZhong_57127
YiChang_57461
WuHan_57494
ChongQing_57516
ChangSha_57679
ZhiJiang_57745
GuiYang_57816
XuZhou_58027
QingJiang_58144
NanJing_58238
ShangHai_58367
AnQing_58424
HangZhou_58457
NanChang_58606
QuZhou_58633
WenZhou_58659
FuZhou_58847
YongAn_58921
LiuZhou_59046
XiaMen_59134
WuZhou_59265
GuangZhou_59287
ShanTou_59316
NanNing_59431
ZhanJiang_59658
HaiKou_59758

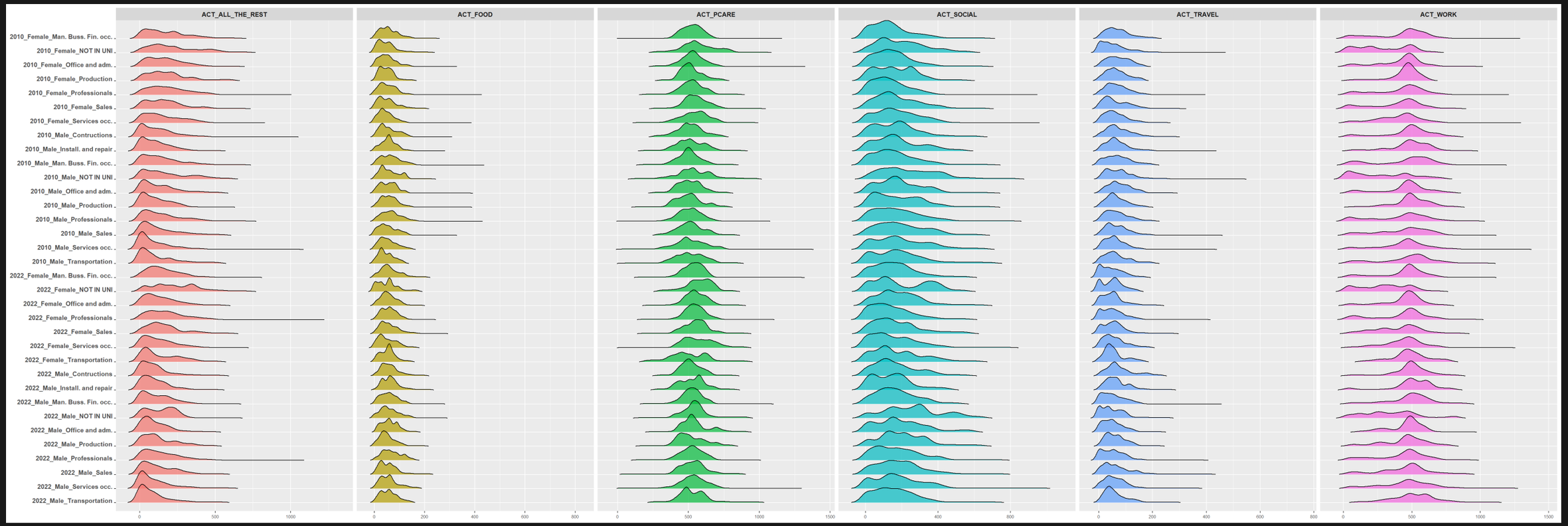
The OZONE dataset 78 obs, 4 vars





Other distributional datasets

- HMAT Time use distributional data created from from an IPUMS dataset available from https://github.com/Airpino/Symbolic_data_analysis_softw/raw/main/Histo_data_IPUMS.RData



Some exploratory tools

Some links showing some exploration tools for distributional data

- An application on climatic data: https://airpino.github.io/ICDS_23_presentation/#/title-slide
- An application on EUSILC 2013 data: https://airpino.github.io/DSSR2024_pres/#/title-slide

An analysis task

1. Choose a dataset
2. Explore the data
    - compute some basic statistics from data
3. Fix some objectives
4. Analyze the data (or a selection of the data) using at least one of the following techniques
    - Principal components analysis
    - Regression analysis
    - Clustering analysis
5. Present the results
6. Let’s discuss together