RFA 2019 - Databases

UNIVERSITÉ DE NEUCHÂTEL

Loïc Maréchal

October 31, 2019

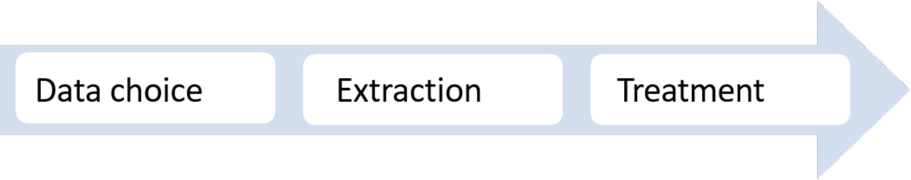
Objectives

- Understand the data collection process
- Learn which data is available from each provider
- Become efficient and autonomous

Outline

- The data collection process
- Thomson Reuters Datastream (Eikon)
- CRSP-Compustat
- Technicalities
- Other available databases
- Application

Data collection



- What data do I need ?
- From which / how many sources ?
- How do I match them ? (in cross-section and time series)

Data collection - variables

- What are the variables of interest ?
- For which dimensions (firms, countries, time periods etc.)
- Where / how to efficiently download them
- How to match them, compute them from raw data
- In brief, as for programming languages, write pseudo-code of the process

Data collection - sources

- What databases do I need ?
- Does the databases subscription covers the data I need ?
- Always refer to the database user guides for data description and availability

Data collection - multiple sources and matching

- Need to match the data within a single database
  - Key identifiers
  - Filters
- Need to match the data within several databases
  - What are the matching keys ?
  - What is the extra information to compute matching keys ?

Data collection - example of time series matching

- Prices sampled at high frequency (one or five minutes) to compute daily measures of RV
- Requires precise matching with daily closing prices and additional matching with weekly and monthly time series
- Matching precision is essential, biases arising from leading information shows up very quickly (type 1 error), type 2 error also exists when too conservative lags are added
- Example 1- Non synchronicity of daily data from different exchanges the NIKKEI and S&P 500 futures
- Example 2- Forward looking information effects on backtesting

Data collection - cross-sectional matching

- The matching is similar to that of time series, as long as the keys are well defined
- The keys for stocks (identifiers) may be : CUSIP, PERMNO, tickers, company name, etc.
- For futures contracts and indices it is more common to use the exchange tickers

Data collection - interrogation process

- Choice between static or time series information
  - Static = m variables, n companies, one date
  - Time series (panel) = m variables, n companies, t dates
- Two main type of variables
  - Market level information (prices, returns, trading volume, OHLC)
  - Company level information (balance sheet, income statements, ratios)

Data collection - interrogation process

- Time span selection
  - Date or date range
  - Time interval (frequency)
- Output format
  - Wide or long form
  - File type (*.csv, *.tsv, *.xlsx)
- In all cases, **do not forget to save the query !**

Eikon (Datastream, IBES, INVESTEXT)

- Datastream
  - Coverage : worldwide listed stocks, futures, options, swaps (partial)
  - Date range : various to "today"
  - Frequency : minute, daily, weekly, monthly, yearly
  - Market level items : prices, returns, volumes, shares outstanding, open-interest
  - Company level items : balance sheets, income statements, ratios
- Investext
  - Analysts reports and data
- I/B/E/S
  - Analysts forecasts (earnings, price targets)
  - Analysts recommendations (buy/hold/sell)

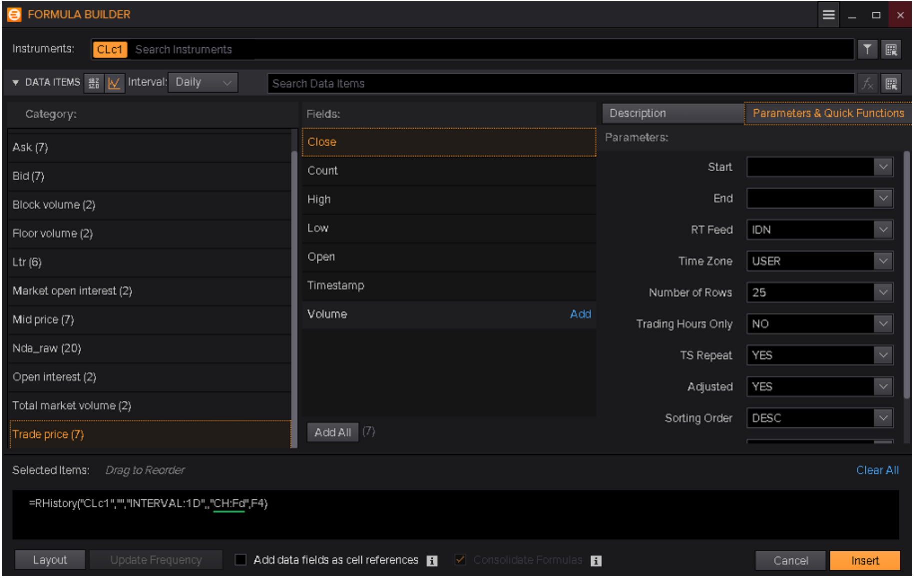
Eikon - other worldwide data

- Bonds
- Unit trusts / investment trusts
- Exchange rates
- Interest rates
- Commodities
- Derivatives
- Economics

Eikon - misc.

- Accessibility
  - SECO library
  - R.103
  - Your own laptop (web or software version) **Coordination with your class is key !**
- Output format
  - To Excel sheet
  - Datastream excel template
  - Eikon live data refresh
- Advantages
  - International coverage
  - Many type of data
  - Flexibility, easy database update
- Drawbacks
  - Only four user logins to share for the institute
  - Remote database - execution time depends on the time of the day
  - Finding the tickers or the appropriate variable name might be complicated

Eikon - demonstration



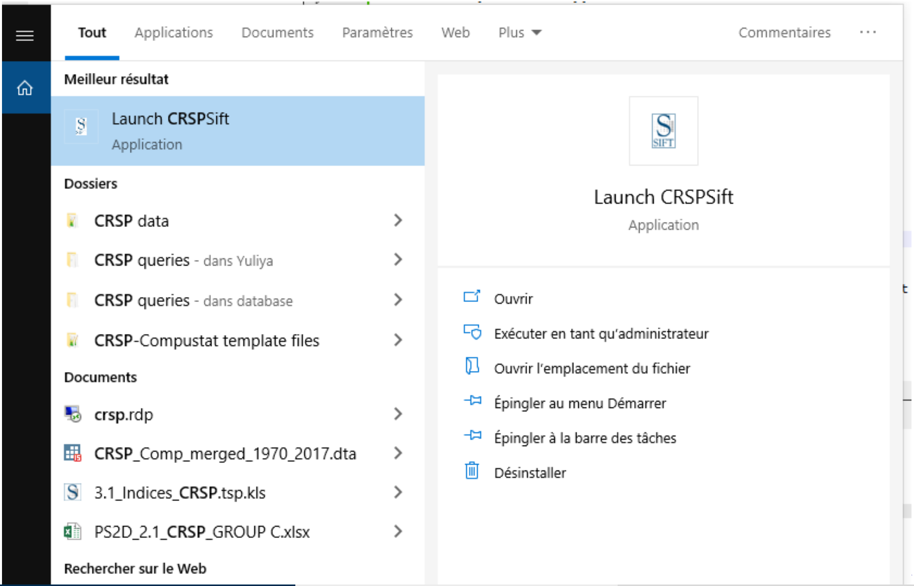
CRSP - COMPUSTAT

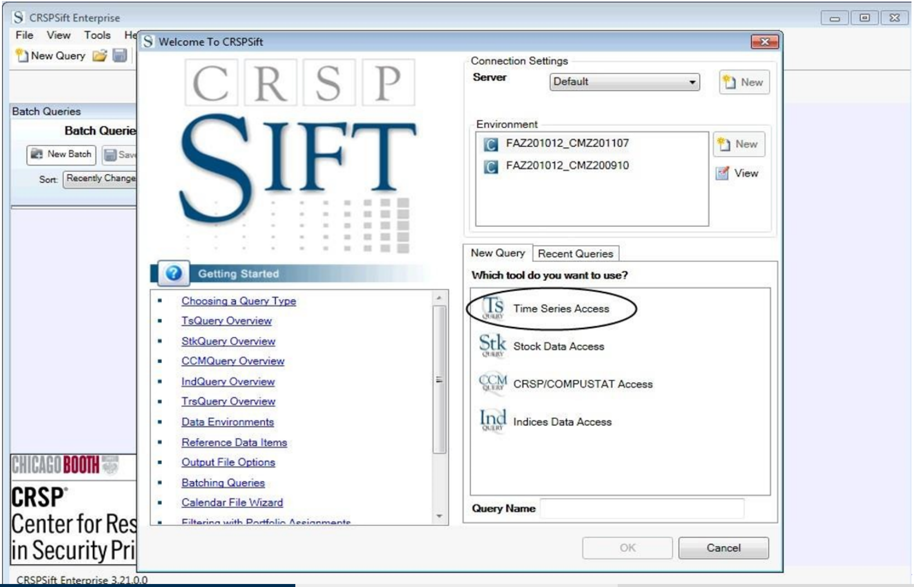
- CRSP : Stock market data
  - Coverage : US listed stocks only
  - Date range : 1927 to December 2018
  - Frequency : daily, weekly, monthly, yearly
- COMPUSTAT : Accounting data
  - Coverage : US listed companies
  - 1949 to December 2018
  - Frequency : quarterly or yearly (depends on the variable)
  - Variables : balance sheet, income statement, sales, amortizations, acquisitions, current assets, debts

CRSP - COMPUSTAT

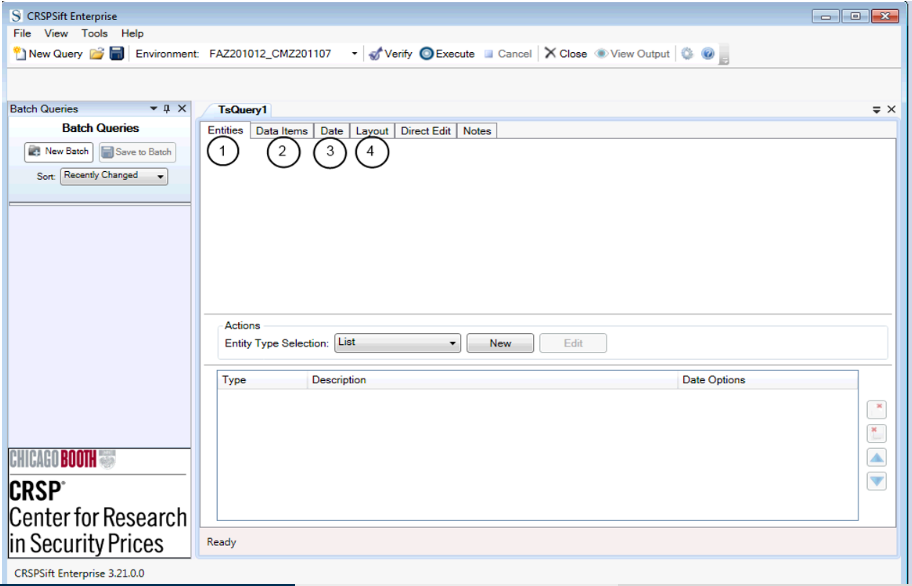
- Accessibility
  - R.103
  - Only through CRSPSift software
- Output format
  - *.txt (recommended) for large downloads
  - *.xlsx and softwares' proprietary formats
- Advantages
  - Extensive coverage
  - Easy to find variables and stock names
- Drawbacks
  - Old software
  - Execution of large queries might fail, need to split them

CRSP - COMPUSTAT




CRSPSift Enterprise
File View Tools He
New Query
Batch Queries
Batch Querie
New Batch
Sort: Recently Change
CHICAGO BOOTH
CRSP
Center for Res
in Security Pri
CRSPSift Enterprise 3.21.0.0
Welcome To CRSPSift
CRSP
SIFT
Getting Started
Choosing a Query Type
TsQuery Overview
StkQuery Overview
CCMQuery Overview
IndQuery Overview
TrsQuery Overview
Data Environments
Reference Data Items
Output File Options
Batching Queries
Calendar File Wizard
Connection Settings
Server
Default
New
Environment
FAZ201012_CMZ201107
FAZ201012_CMZ200910
New
View
New Query
Recent Queries
Which tool do you want to use?
Time Series Access
Stock Data Access
CRSP/COMPUSTAT Access
Indices Data Access
Query Name
OK
Cancel

CRSP - COMPUSTAT



- Selection of one to *n* companies
- Variables (items) selection
- Date range and frequency selection
- Output file formatting

Technicalities - user guides

- See all the documents in the Moodle folder
- Eikon - Datastream
  - Datastream advance for Office - Using Excel
  - Datastream for Office - User manual for Excel
  - Worldscope database - Variable definitions
- CRSP - COMPUSTAT
  - CRSP variable definitions
  - CRSP - COMPUSTAT merged database guide
  - CRSPSift user guide

Technicalities - how to connect

- Eikon - Datastream
  - ID : "Eikon2@unine.ch"
  - PW : "IAFbreguet1"
  - Available on request for the equity research contest : Investext login
  - Same login to connect on the web and to download the software on your laptops
- CRSP - COMPUSTAT
  - Nothing more than launching the software
  - Use : "deleteUserTempFiles.bat" script in the "D" drive if your query does not return anything
- For any additional troubleshooting, contact me : loic.marechal@unine.ch

Other databases

- I/B/E/S with SQL developer
- Barchart minute data on futures and options written on stocks, indices, commodities, exchange rates, interest rates, crypto-currencies
- SDC (mergers and acquisitions data)
- Hedge funds data with HFR, Morningstar, Tass, Barclay and Eureka (up until 2016)
- https ://www.quandl.com/ for free data online, and many more online providers

Application

- Get the daily adjusted stock prices of Amazon for the period 2015–2016
- Get the monthly returns and excess returns (on a value weighted index, VWI) of Walt Disney since inception. Add the VWI to your query and compute yourself the excess returns and compare
- Get the monthly returns on Exxon Mobil Corp. stocks from 2000 to 2018 and their matching yearly current ratio (current assets – total / current liabilities – total)

- “The January effect is a seasonal increase in stock prices during the month of January. Analysts generally attribute this rally to an increase in buying, which follows the drop in price that typically happens in December when investors, engaging in tax-loss harvesting to offset realized capital gains, prompt a sell-off” (Investopedia). Create a long run (1970-2017) systematic strategy that could benefit from this market anomaly. Create the associated calendar file and test it on the CRSP VWI

loic.marechal@unine.ch