Meiqing Zhang
Rodrigo Valdes
Erin M. Ochoa

CAPP 30122
Project Proposal
2017/01/24

Predicting Commodities Prices and Volumes: A Data-Mining Approach

Introduction

Various geophysical, social, and economic factors contribute to commodities prices and volumes in futures exchanges. By aggregating data on a range of factors from a variety of sources, we aim to train a model that predicts prices and volumes for a selection of commodities. We will do this by data-mining: combining data from disparate sources and permuting models in order to find the most robust predictors of prices and volumes for each commodity.

Methods

First, we will need to select the outcome and explanatory variables and their sources. After inspecting all the sources, we will implement tools to crawl and scrape each source in order to compile the necessary data for each variable. Meanwhile, we will need to download sample data and use it to design a data parser for each source.

We will design and implement a database to house the data and perform aggregate calculations. For example, stock-market data may be daily, but we will apply transformations to such data in order to construct monthly measures. Once the data are in their final form, we will join them based on month and year.

We must decide on an analysis approach. We will use linear regression with either best-k variables using forward selection or principal component analysis. The former is easier to interpret, but the latter can account for complex relationships between explanatory variables; in adopting a method, we will inevitably sacrifice one feature for the other.

Having chosen an analysis method, we will train a model for each commodity of interest using historical data. We will test each model on recent data to determine not only the best predictors for each commodity but also which of the commodities are most robustly predicted.

It is our goal to design and implement functions to automate data visualization. Such algorithms will take in data from explanatory and/or outcome variables and create appropriate visualizations (likely using matplotlib).

Data Sources

Because we are taking a data-mining approach to model design, we will need data from a selection of sources. For example, we may choose to use any or all of the following variables and associated sources:

- Commodities prices and volumes (https://finance.yahoo.com/commodities)
- Stock indices (https://finance.yahoo.com/world-indices)
- Currency exchange rates and/or volumes (http://www.x-rates.com/historical)
- Inflation in the United States, China, Japan, Germany, the United Kingdom, India, and Brazil (https://data.oecd.org)
- Electricity usage in the United States (http://www.eia.gov/electricity/monthly)
- Number of months elapsed since the beginning of the most recent American economic recession (http://www.nber.org/cycles.html)
- Global mean temperatures (http://berkeleyearth.org or https://data.giss.nasa.gov/gistemp)
- Global total precipitation (http://neo.sci.gsfc.nasa.gov)
- Oil prices (http://www.indexmundi.com/Commodities/?commodity=crude-oil)

Tasks and Timeline

What follows is a list of the tasks required to produce this software and an accompanying presentation, along with an anticipated completion week for each task.

Task	Anticipated Completion Date
Assign tasks to group members	4th week
Define selection criteria for commodities	4th week
Gather and analyze data for commodities selection	5th week
Select 20 commodities of interest	5th week
Finalize selection of explanatory variables and data sources	4th week
Download sample data for each variable	5th week
Code crawling and scraping tool	6th week
Code data-parsing tool	7th week
Crawl and scrape all data necessary	8th week
Parse all scraped data	8th week
Design database schema using sample data	5th week
Code database framework using designed schema and sample data	6th week
Populate database using parsed data	8th week
Calculate aggregate transformations on daily data to result in monthly values	8th week
Decide on analysis method (best-k or principal component analysis)	4th week
Design analysis algorithm	5th week
Code analysis algorithm	6th week
Train models using historical data	9th week
Test models on recent data	9th week
Design visualization functions	5th week
Code visualization functions	7th week
Process visualization functions	9th week
Begin documentation	8th week
Complete documentation	9th week
Begin final presentation	9th week
Integrate visualizations into final presentation	9th week
Complete final presentation	10th week
Present project	10th week
Finalize software	03/14