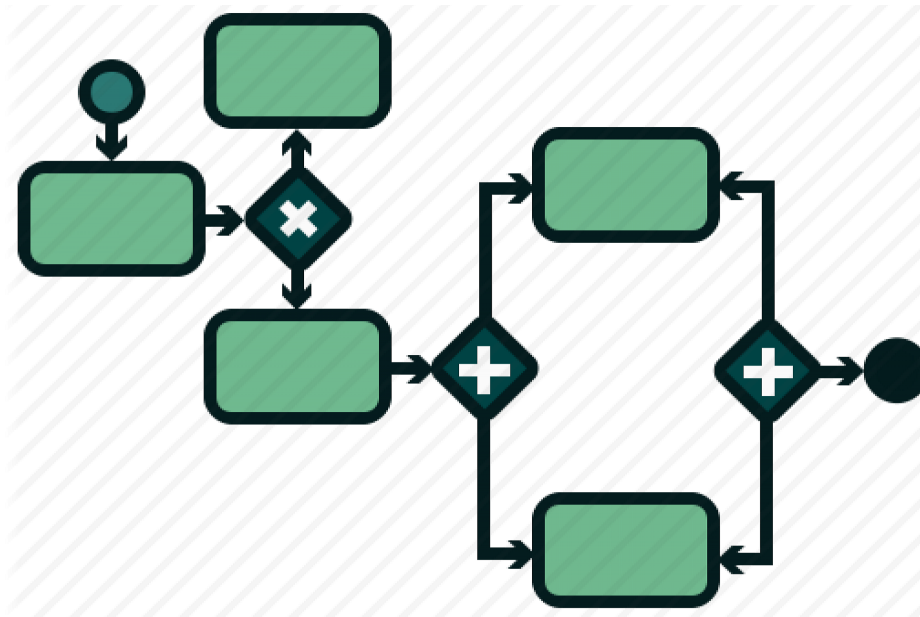


EPSRC Centre for Doctoral Training in Industrially Focused Mathematical Modelling



Cleansing of Financial Market Data: A Generic Framework

Victor (Sheng) Wang



Contents

1 Introduction	1
Motivation and Goal	1
Approach	1
Glossary of Terms	1
2 Financial Pricing Unit	2
Financial Pricing Unit Construction	2
3 Principles, Framework, and Processes	4
The Framework	4
Processes	4
4 Discussion, conclusions & recommendations	6

1. Introduction

Motivation and Goal

CME Clearing (referred to as “CME”) use a historical simulation approach to generate risk scenarios for the portfolios of financial products of their clients. The distribution of profits and losses is constructed by taking the current portfolio and subjecting it to the actual changes in the risk factors experienced during each day of a historical period. To simulate reasonable risk scenarios, it is crucial to ensure the historical data does not have missing values, outliers or invalid prices before entering into the simulation engine. The process of addressing these issues is called *data cleansing*.

We aim to construct a generic framework that is capable of consistently describing the mainstream methods and processes that are currently in use at CME to cleanse financial data associated with a variety of products and asset classes. The framework should provide quick guidance on the particular workflow to follow and an encyclopedia of techniques to use during a data cleansing task. This will benefit CME (1) by accelerating risk modelling research and promoting fast launch of new products, and (2) by comparing different methods and processes in the framework so that CME can align their methodology across the global team.

Approach

We start by defining the concept of a *financial pricing unit* on which the data cleansing task is carried out. A financial pricing unit (referred to as “FPU”) is a structure of a financial variable used as an input by a pricer (pricing model) to evaluate any financial product.

The key principles that govern the properties the cleansed data should have are (1) completeness, (2) outlier-free, and, (3) no violation of pricing constraints. These principles give rise to three corresponding processes to cleanse data, namely data completion, outlier detection and price validation. These processes are organised in a particular order to form the generic framework. There are possibly multiple approaches to accomplish a process. We call each approach a method. The relationship between framework, process and method is demonstrated in Figure 1.

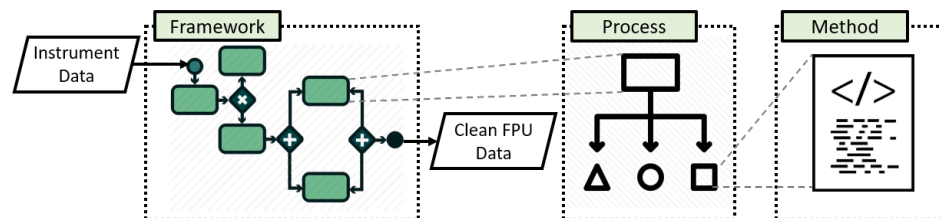


Figure 1 – Data cleansing framework, process and method.

Glossary of Terms

- **FPU:** Financial pricing unit (FPU) is a structure of a financial variable (eg. interest rate, exchange rate) used as an input by a pricer to evaluate any financial product.
- **Product feature:** A financial derivative product is characterised by a set of product features such as an expiration/maturity date (the last day of validity), moneyness (the intrinsic value in its current state), etc.
- **Data skeleton:** Data are only collected for a set of traded products with more reliable prices. The product features associated with this set of products form a data skeleton.
- **Data completion:** A process to fill any value absent on the data skeleton.
- **Outlier detection:** A process to identify rare observations which raise suspicions by differing significantly from the majority of the data.

Financial data are clean if there is no missing value, outlier or invalid price. Data cleanness is crucial for a good risk modelling.

- **Price validation:** A process to check whether a data value incurs significant violation against efficient market asset pricing theories or market conventions.

2. Financial Pricing Unit

With generic data cleansing framework, we aim to cover some specific products of specific asset classes. In Table 1, we present the asset classes and products under consideration by CME.

Asset Class	Product	Features
Rates	Zero coupon swap, Basis swap, Forward rate agreement (FRA), Overnight index swap (OIS)	Variable time-to-maturities; Variable currencies; Variable fixed cashflow frequencies (1-day, 1-month, 3-month and 6-month).
	Interest rate swap (IRS)	Time-to-maturity up to 50 years, for each denomination currency; Cashflow frequencies include 1-month, 3-month, 6-month, subject to the denomination currency; 24 denomination currencies.
	Interest rate swaption	Time-to-expiry up to 2 years; Underlying maturity up to 30 years; Cashflow frequencies include 3-month only; Any strike; 1 currency: USD; Physically-settled vanilla European options.
FX (Foreign eXchange)	FX forward	Time-to-maturity up to 2 years; 26 cash-settled forwards; 11 non-deliverable forwards.
	FX option	Time-to-expiry up to 2 years; Any strike price; 7 (out of G10) currencies; Cash-settled vanilla European options.
Commodity (Energy)	Crude oil futures and options	There are 108 futures (outright, spread) and 104 options (outright, spread) listed. Major futures: CL - Crude oil futures, BZ - Brent last day financial futures; Major options: LO - Crude oil option, LC - Light sweet crude oil European financial option;
	Refined products futures and options	There are 332 futures (outright, spread, crack spread) and 47 options (outright, spread, crack spread) listed. Major futures: RB - RBOB gasoline futures, NY harbor ULSD futures; Major options: OH - NY harbor ULSD option, OB - RBOB gasoline options;
	Natural gas futures and options	There are 53 futures (outright, basis, index) and 65 options (outright) listed. Major futures: NG - Henry hub natural gas; Major option: LN - Natural gas option (European), ON - Natural gas option (American).

The generic data cleansing framework is scalable to various asset classes and products.

Table 1 – Asset classes and products supported by CME risk management services capacities.

Financial Pricing Unit Construction

A typical pricing model assumes that the price of a financial product is a function of several financial variables and product features. For example, to price a European vanilla call option using the Black-Scholes model, we need not only the option strike price and expiry from the product features, but also financial variables including the risk-free interest rate, the price of the underlying asset, and the volatility of the underlying asset.

We define a *financial pricing unit* (FPU) as the structure of a financial variable used as input by a pricing model to evaluate any financial product under consideration. The structure of a financial variable consists of:

- the multi-dimensional nature characterised by the product features. The market data of a financial variable are jointly determined by product features such as expiration date, moneyness, etc. The number of dimensions of the financial data is one larger than the number of product features because we also have to include time.
- the skeleton where market data for traded products are collected. Data are only collected for a set of well traded (liquid) products with more reliable and informative price.
- the interpolation and extrapolation scheme performed on the data collected on the skeleton.

To price an arbitrary financial product, we interpolate and extrapolate on observed discrete values of the financial variables.

In Figure 2 we show the FPUs required for pricing all the products under consideration.

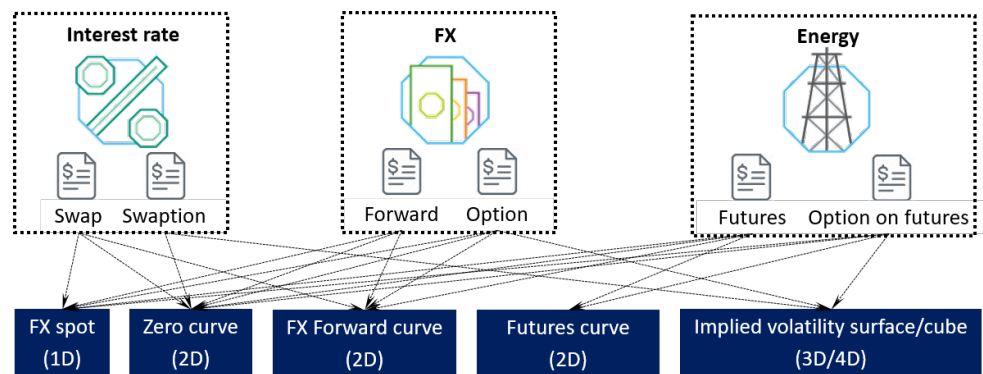


Figure 2 – Map of financial pricing units required for pricing different products in scope.

By defining an FPU, we solve two difficulties in looking for the value of financial variables at an arbitrary product feature. First, many financial variables are either not directly observable or observable from thinly-traded instruments, in which case their values need be implied from market quotes of liquid instruments. Second, we only have discrete financial variable data on the data skeleton. Therefore, interpolation and extrapolation techniques are needed to construct continuously-valued financial variables.

The construction of an FPU consists of three steps: (1) select a set of actively traded financial instruments on the pre-defined data skeleton, (2) build the FPU skeleton by implying data from the market quotes of the selected instruments (a inverse pricing problem), and (3) build continuity on the FPU skeleton through proper interpolation and extrapolation scheme. An example of constructing an FX forward curve is given in Figure 3.

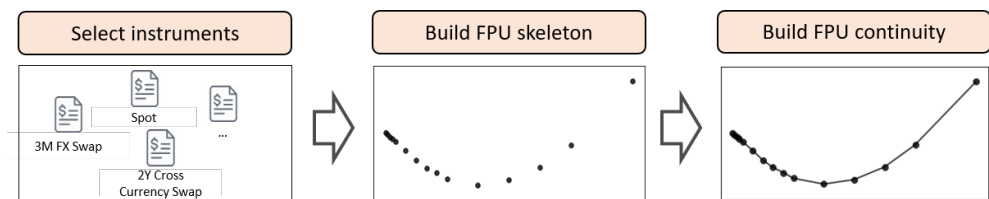


Figure 3 – An example of constructing an FPU, FX forward curve.

The market data for FPUs varies over time and has multi-dimensional nature. In particular, we refer to 1D data as time series, 2D data as term structures or curves, 3D data as surfaces, and 4D data as cubes. In Figure 4 we visualise some example data of different dimensions.

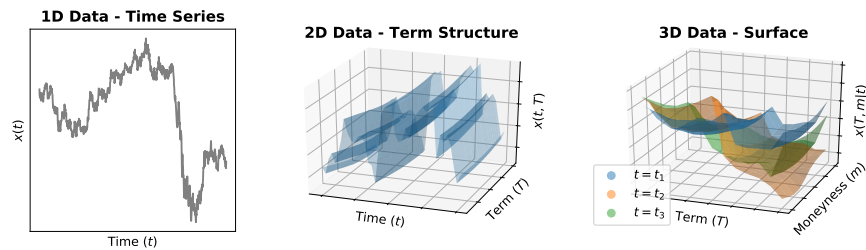


Figure 4 – Visualisation of representations of 1D, 2D and 3D data.

3. Principles, Framework, and Processes

When constructing the FPU following the steps outlined in Figure 3, we need ensure that the data are “clean”, otherwise data cleansing processes need to be applied. Data cleansing *principles* are high-level descriptive guidelines to follow and requirements to fulfil. We formulate the following principles by considering the pre-requisites of the downstream historical risk scenario simulator, respecting the fundamental theorem of asset pricing under the Efficient Market Hypothesis, and appreciating asset-specific market conventions.

1. **Completeness:** during the construction of an FPU, market data are collected on the data skeleton (the pre-defined set of times, maturities/expiries, and monynesses). Data are said to be incomplete if any numerical value is absent on the skeleton. If we have incomplete data, the risk scenario simulation will not proceed.
2. **Outlier-free:** unidentified outliers in the data set can introduce spuriously extreme events to the risk scenario pool. The risk model is not robust to outliers so the computed risk will be overestimated in the presence of outliers. Data cleansing should detect all outliers and identify the cause.
3. **No violation to pricing constraints:** the Fundamental Theorem of Asset Pricing under the Efficient Market Hypothesis infers arbitrage-free conditions on a single FPU, or arbitrage-free relationships among multiple FPUs. Data cleansing should be able to detect non-trivial arbitrage opportunities implied from data and identify the cause.
4. **Respecting market constraints:** the market constraints are the constraints that are subject to the nature of the asset class or the market where the financial products are traded. Examples include the bid and ask boundaries, tick size restriction on the price movement, and macroeconomic event that supports or disapproves the observed values. Some market constraints are often difficult to use due to the limited availability of data.

The Framework

Our principles give rise to multiple *processes*, each of which is designed to fulfil one specific requirement, such as to detect outliers, to detect violations to pricing constraints. These processes are organised in a particular order to form a *framework*. We propose a generic framework, shown in Figure 5, that is capable of describing all existing data cleansing workflows used in different asset classes.

The framework is FPU-orientated, since it starts with instrument selection and ends with building continuity of the constructed FPU. Other processes serve as intermediary steps that ensure satisfactory data quality. These will become optional if the data is already clean.

Processes

The process “validate price” is proposed in accordance with Principles 3 and 4. It appears twice in the framework, once before the FPU is constructed and once after. Normally we can detect violations of pricing and market constraints from the instrument data, but there are cases where the constraints are not explicit for instrument data. For example, there is no established arbitrage-free condition for swap rates. Therefore the second validation process on the FPU becomes necessary.

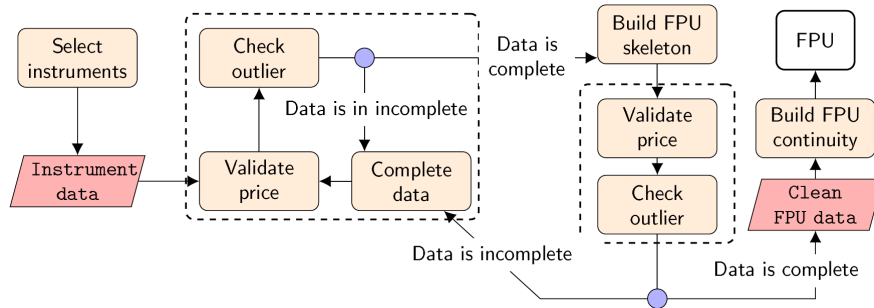


Figure 5 – A generic framework of data cleansing processes.

Principle 2 leads to the process “check outlier”. An outlier may indicate errors in either data recording or measurements, or may be due to rare or unexpected macroeconomic events. We will remove erroneous outliers but keep market-event-supported outliers.

Invalid price or outlier will be removed and then be filled in the data completion process.

A data point that either violates market or pricing constraints, or is thought to be a suspicious outlier, will be removed and treated the same as a missing value. It will be replaced by a new value, in the same way as a missing value in the raw data set is filled, except that the original value might provide useful information for the replacement. The process of replacing a bad value or filling a missing value is called “complete data”, in response to Principle 1.

We show an example of various data issues and their corresponding data cleansing processes in Figure 6.

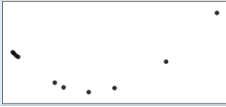
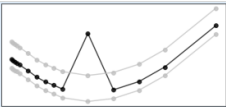
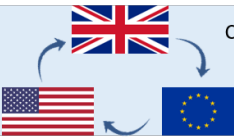
Principles	Data Issue	Process
Completeness	 <ul style="list-style-type: none"> Data are missing on the skeleton 	Complete Data
Outlier-free	 <ul style="list-style-type: none"> There is a kink on the curve The curve is inconsistent with 	Check outlier
No violation to pricing constraints	 <p>Currency triangular relationship:</p> $F_t^{EURGBP}(T) \neq \frac{F_t^{EURUSD}(T)}{F_t^{GBPUSD}(T)}$	Validate price

Figure 6 – An example of three processes for data cleansing and there corresponding principles.

We perform the “check outlier” and “validate price” processes twice, once on the instrument data, and once on the FPU skeleton data. This not only acts as a double check but also become necessary in situations where constraints can only be imposed on either the instrument data or the skeleton data but not both. The “complete data” process is nevertheless only performed on the instrument data, because we need understand the change in source information. One data point on the FPU skeleton might be influenced by multiple instrument data points. Directly filling an FPU data point without understanding how the underlying instrument data varies might result in inconsistencies between data points.

4. Discussion, conclusions & recommendations

We have defined the principles for cleansing financial data and have formulated a generic and consistent framework that provides quick guidance on the workflow to follow, and an encyclopedia of techniques to use during a data cleansing task. The framework covers processes and methods that have been successfully applied in cleansing tasks for multiple asset classes (rates, FX, energy) and products (swap, forward, futures, options) within the scope of CME's risk management capabilities.

All data cleansing tasks will have to go through three processes to get the clean financial pricing unit (FPU) data. The three processes we propose are price validation, outlier detection, and data completion. Price validation is the most generic process in which very similar methods are used across products and asset classes to detect the violations against pricing theory and market constraints. There is limited genericness in both outlier detection and data completion processes.

Our work has focused on a comprehensive review of the mainstream data cleansing processes and methods used at CME, and summarises similarities and differences. Further analytics should be undertaken to evaluate and compare the performances of different methods. There is room to improve current methods. For example, for the sake of data completeness, interpolation and extrapolation are needed when there are missing values, outliers, arbitrageable values, etc. Current techniques mainly focus on geometrical smoothness criteria without considering the underlying financial process. Consistent interpolation can be built by modelling the generative process of financial data and the interconnectedness among the financial variables. Consistent interpolation will propose a new value for the financial variable that has the maximum likelihood given a model and observed data, respects arbitrage-free pricing, and satisfies smoothness requirement.

Florian Huchedé, Director of Quantitative Risk Management at CME Group said: *"We are very pleased to have Victor work on this comprehensive inventory and review of data cleansing practices in use across our global quant teams. His work will serve as a coherent part of our ongoing corporate-level risk model and methodology alignment project. He has built the generic framework for data cleansing. This will benefit CME by accelerating risk modelling research and promoting fast launch of new products. Discussions are already underway to form a global data cleansing team that could leverage Victor's proposed framework to build technical infrastructure for a generic data cleansing solution. We are looking forward to working with Victor in the upcoming long-term DPhil research project to have deep dive on the topic."*