U.K. cross-sectional equity data: do not trust the dataset! The case for robust investability filters

Francesco Rossi

July 2011

U.K. cross-sectional equity data: do not trust the dataset! The case for robust investability filters

Francesco Rossi

UCD Michael Smurfit Graduate Business School

Banking & Finance

Dublin, Ireland

francescorossi_@hotmail.com

Abstract

We propose a novel approach to cross-sectional equities sample selection, derived from best market practice in index construction and focused on investability. Using the U.K. market as a template, we first demonstrate how the popular Datastream dataset is plagued by data deficiencies that would surely invalidate statistical inferences, and that are not addressed by commonly used filters. We show the benefits and need for a supplementary data source. We then develop robust investability filters to ensure statistical results from cross-sectional analysis are economically meaningful, an issue overlooked by most studies on cross-sectional risk pricing

DUBLIN, IRELAND

NOVEMBER 2011

Contents

1. Introduction

The literature on cross-sectional equity pricing has grown significantly in recent years, expanding and refining earlier analysis with respect to the variety of cross-sectional effects tested and to the coverage of international markets in addition to the US. This has been in part driven by new cross-sectional equity datasources becoming available to researchers. International markets datasets offer a more limited set of cross-sectional details, encompassing markets with few stocks, sparse trading, and, especially prior to the last two decades, they have been much less scrutinized, with problems related to market microstructure and data quality becoming more relevant. However, despite the obvious relevance of ensuring a high-quality of such data for accurate and relevant analysis, we have found researchers to be generally very little aware of this issue, resulting in very low standards when it comes to implementing data-quality checks. Moreover, a paper by Ince and Porter (2006) highlighted in vivid details how serious deficiencies in the most commonly used dataset for European cross-sectional equity data, Thompson Datastream, so that a naïve use of such dataset would make economic inferences totally unreliable. Yet, when we reviewed older and newer literature on cross-sectional equity risk pricing, we found a generalized and serious underestimation of this issue, with the data-quality filters used being very weak. No cross-checks against alterntive data sources are usually made (and we will show the importance of doing this). Perhaps not surprisingly then, this very same stream of literature seems to be endlessly debating on the relevance of residual risk as a priced risk factor, with opposite and conflicting results. We will indeed show that the data is so bad and flawed that statistical inferences driven without a thorough review and correction exercise are, at best, totally unreliable.

2. Contribution

We add to the literature looking on cross-sectional equity analysis, by means of a comprehensive review of the Datastream U.K. sample, documenting serious issues of data quality, and dealing with them by means of a rigorous review and of an innovative methodology based on both data-quality and investability filters. We propose a novel approach to sample selection derived from best market practice in index construction and focused on investability.

Our work focuses on a U.K. dataset, but obtains results with a general scope on both the methodological approach and the empirical results, applicable to ample and diversified pools of stocks, with certain homogeneous characteristics, traded in a single market. We confirm all the troubling findings of Ince and Porter (2006). We first demonstrate the seriousness of the data issues, and the significant effort a researcher must undertake to get close to an “error-free” sample. Unfortunately, we show that this likely to be impossible unless one can avail of a second data source to cross-check Datastream information. In other words, Datastream should be avoided if possible. Secondly, we focus on one issue closely related to data-quality but overlooked by most studies on cross-sectional risk pricing: how to ensure that cross-sectional equity data are economically relevant, i.e. the constituents of the sample are investable. We develop robust investability filters to ensure statistical results from cross-sectional analysis are economically meaningful and relevant for institutional market participants, using an approach derived from best market practice in index construction.

3. Structure of the paper

Section 4 starts with a description of our dataset, then deals with data quality and the process of data correction and data quality filters. We build and expand on a survey of previous literature on cross-sectional equity risk measures, and we demonstrate the seriousness of the data deficiencies in the Datastream sample. In Section 5 we deal with the issue of investability and liquidity, developing additional filters to ensure investability and liquidity, derived from market practice in index contruction, and apply them to our sample, showing and commenting on the results. Section 6 presents the conclusion of our analysis.

4. Dataset and data methodology

4.1 Data Sourcing

We base our initial data query from Datastream. The starting universe includes all stocks listed on the London stock exchange since 1990. We also use an alternative data source, Bloomberg, in order to deal with data issues identified in the Datastream raw sample; to do so, we match securities by ISIN and SEDOL codes to fill-in missing values, cross-check dubious values and replace erroneous data. In some cases, we resort to company websites to obtain historical information related to outstanding shares.

We collect the following cross-sectional data from 31/01/1990 to 31/12/2009, at monthly frequency:

i. Price
ii. Return Index (a total return index that accounts for corporate actions such as, for instance, dividend distributions and stock splits)
iii. Volume
iv. Number of Shares outstanding

In addition, we have obtained sector classification data with 4 levels.

The selected classification standard has been the Industry Classification Benchmark “ICB” hierarchy. This is a standard classification system developed by FTSE Group and Dow Jones

Indexes, managed on a transparent and rules-driven basis. It contains four classification levels: Industries (10), Super-sectors (19), Sectors (41), and Subsectors (114). Allocation to the different codes is driven by revenue not profit, and this applies to industry as well as country classification. For instance regardless of the listing exchange, the sector index identified for each stock is the one of the country where the company generates the largest revenue share.

We also collect the series of 1-month GBP Libor rates from Bloomberg as a measure of "riskless" monthly interest rate.

4.2 Data Quality issues in previous literature

We want to define filters to ensure good sample data quality. We start from a review of the practices commonly used in the literature on cross-sectional risk.

It is possible to separate three key dimensions of the data quality problem:

- choice of data provider and perimeter of sample universe
- assessment of the quality of recorded data
- methods to reduce the impact of data errors

The latter two are obviously closely linked.

Most of the existing research on idiosyncratic risk uses U.S. data. The dominant data source for U.S. cross-sectional data is CRSP, complemented by Compustat for financial accounting metrics. On the other hand, research on international data, carried out by fewer researchers, has often applied the same framework but resorted to different data sources, most frequently Datastream, with some exceptions for the UK (see Li et al. (2008)) and Japan where alternative sources are available.

We have surveyed sample selection criteria used by U.S. researchers, spanning the years from 1962 to roughly the mid of the last decade. We report them hereafter, although regrettably in some cases the provided disclosure on the methods used falls short of being accurate.

Using CRSP codes 10 and 11, most restrict the sample to common shares of U.S.-incorporated companies. In terms of choice of sample universe, some researcher highlights that using stocks trading on different exchanges could introduce a bias due to heterogeneous market structure, and filter out non-NYSE stocks (AMEX/NASDAQ); we think this is a sensible decision. A minimum number of data points is needed for model estimation, and some authors give detail on their chosen minimum: Spiegel and Wang (2005) require return, shares, price, volume for at least 24 out of the past 60 months, while Arena et al. (2008) require 12 months of return data and Fu (2009) requires 30 months. Fama and French (2008) exclude firms with missing details on market cap or accruals.

Moving to strictly data-quality checks, Bali et al. (2005) require "valid" (?) prices. Cremers and Mei (2007) require "stable" turnover and exclude stocks with volatility greater than 10 times the average volatility. Some authors are concerned with missing returns[1]: Campbell et al. (2001) require returns for at least 75% of the days in the past year, while Fama and French (2008) need the return of the previous month to compute lagged 12-months returns, and Garcia and Martellini (2009) use some form of missing returns filter.

Price level is also a concern. Bali et al. (2005) highlight that returns on low-price stocks are greatly affected by minimum tick rules imposed by exchanges, which may add noise to the construction of return and risk measures: most exchanges require that quotes and transaction prices be stated as some multiple of a minimum price variation, or trading tick. Consistently

[1] We treat the case of missing return data distinctly from the case of zero returns, that we review in the following section dedicated to liquidity and investability

with **Jegadeesh and Titman (2001)** they thus choose to exclude stocks priced below $5, selecting the threshold that was also used by NYSE in 1992 when it reviewed tick rules allowing stocks priced under $5 to trade with smaller ticks for smoother pricing. They are followed by Arena et al. (2008) and Ruan et al. (2010) who also exclude stocks trading below 5$ so that the discreteness of price movements below this level does not bias returns[2]. Extra-large jumps in prices or returns are likely to be the product of data errors, thus Campbell et al. (2001) and Fu (2009) employ a filter on extreme price movements, excluding stocks with daily or monthly returns in excess of 200% or 300%, respectively. Fu (2009) also suggests using log returns and GLS techniques to mitigate the impact of data errors.

Researchers using international data face extra challenges, due to heterogeneous samples in terms of country attributes, and also due to sample data availability and quality; as a result, their samples tend to start not earlier than the '80s. However, the quality and quantity of filters employed to improve data quality differ markedly even within studies depending on the same data source. Bartram et al. (2009), Brockman et al. (2009) and Guo and Savickas (2008) rely on Datastream for data on international markets. The first two papers restrict the sample to common stocks. Bartram et al. (2009) make also use of Worldscope as a source of country classification, requiring stocks to be denominated in a currency that is legal tender in the listing country and excluding stocks lacking country/stock identifiers or the selected financial reporting data used in their study; they also require a minimum of 25 weeks of consecutive returns, while Brockman et al (2009) require 30 months, and Li et al. (2008) require 3 months for a UK-based sample sourced from LSPD.

---

[2] See Bali (2005) for detailed rationale and background. See also MSCI methodology below for more background. We have incorporated this aspect in our filters

To address some concerns about data errors, Brockman excludes monthly returns greater than 200%, while Bartram et al. (2010) and Guo and Savickas (2008) follow Ince and Porter (2006) implementing a filter for return reversals that could be caused by incorrect stock prices, combining a 300% limit with a reversal alert. Bekaert et al. (2007) measure the proportion of zero-return days and find it to be related to liquidity, that they are trying to proxy due to lack of transaction data. Bartram excludes stocks with zero returns for more that 30% of the weeks in a year. Guo and Savickas (2008) implement several rules to address potential data errors, such as setting a minimum value for the return index and filtering out sharp return reversals.

Ince and Porter (2006) exclude stocks trading below 1$ to avoid issues with discrete price jumps driven by Datastream rounding practice.

In fact, it is the Ince and Porter (2006) paper, significantly echoed in most of the filters used by Bartram et al. (2010), that needs to be closely considered as far as data-related issues are concerned, as it provides ample evidence that the handling Datastream data for U.S. or international samples needs special care. They document how raw data contains a significant amount of incorrect information, both qualitative (classification information) and quantitative (prices, returns, volume, shares etc). Unless proper techniques are used to correct them, inferences drawn from raw or lightly filtered data are, at best, dubious. In particular, one key finding is that within Datastream it is not possible to distinguish easily – i.e. using information provided by Datastream itself - between the various types of securities traded on equity exchanges; for instance, many securities classified as common stocks are not such. There is no easy method to tackle this issue, other than access to a second data source and a careful screening performed on the securities' names, to identify non common stocks securities based on key words. Classification issues do not end there, as they find that the full time series of classification variables often reflect only the most current value. They find issues of coverage and a survivorship bias that may introduce distortions into a sample of comparable firms. They run a comparative analysis with CRSP data, revealing several problems with quantitative data, some of which would be difficult, if not impossible, to isolate without an alternative data source. For example, they find instances of stock splits reflected in the Datastream data on incorrect dates, differences in closing prices and dividend payments. They document several issues with total returns calculation, with the time markers for beginning and ending points of price data and with handling of returns after suspension periods[3]. They also flag problems caused by Datastream rounding of stock prices below 1$ and with small values of the return index. They determine that most (not all) of the problems identified are concentrated in the smaller size deciles, showing how this issue would significantly impact inferences drawn by studies focusing on cross-sectional stock characteristics. On the positive side, they show that implementing the screens and filters they propose[4], the scale of the problem can be effectively reduced, and inferences drawn from corrected Datastream data are very similar to those drawn from the CRSP sample. They show that data on several international markets are heavily affected by the same data issues documented for the U.S. sample. They compute cross-sectional, market capitalization-weighted sample averages of returns for each country, and compute the correlations with the returns of country total market indexes[5]. The correlations are as low as 0.20 before applying any filtering technique, but rise to 0.98 after implementation of their filters, indicating a success in dealing with the bulk of the issues. Interestingly, the UK market seems to be affected in a particularly severe fashion by securities misclassification: unless one introduces filters aimed at removing issues improperly classified as common stocks,

[3] We have documented all of the listed issues in our sample, some of which can be corrected only through a manual, name-by-name and Isin-by-Isin check. Numerous cases of issues classified as equities that were actually unit or investment trusts; obvious data recording errors on the outstanding shares data, where adjustments for corporate action such as stock splits an new issuance was often inaccurate or missing even for large caps and where at times spurious zeros were added (or missed) to the figures; gaps in the volume series; incorrect start and end dates

[4] For instance, they implement a return reversal filter

[5] Total market indexes are market value-weighted, and are disseminated and calculated from independent sources: they can thus be used as proxies for “error-free” value-weighted sample aggregates

data-based filters alone do not bring the correlation to elevated levels. On the other hand, once the classification filters are employed, the correlation jumps to 0.99.

4.3 Our preliminary sample and data quality filters

We queried Datastream for Equities (Instrument type=Equity) listed on the London Stock Exchange, (Exchange=London), and got a potential universe of 7968 securities. The raw data has been carefully reviewed and filtered as follows.

We exclude:

i. Investment trusts and other types of non-common-stock instruments, eliminating securities not flagged as “EQ” (equity) on the “TYPE” data field in Datastream
ii. Securities not denominated in GBP
iii. Unit Trusts, Investment Trusts, Closed-end Funds, Preferred Shares, ADRs, Warrants, Split Issues
iv. Securities without adjusted price history (“ADP” flag in Datastream[6])
v. Securities not flagged as “major securities” (“MAJOR” flag in DS[7])
vi. Securities flagged as secondary listings for the company (“ISINID” not equal to “P”[8])
vii. Stocks identified as non-UK under the ICB classification system[9]
viii. Securities without a minimum return history of 24 months according to start and end date flags provided by Datastream (fields BDATE and TIME)[10]

---

[6] ADP=1 indicates that price history is adjusted for corporate actions, such as share splits and rights issues

[7] For companies with more than one equity Security MAJOR returns Y (yes) or N (no) to indicate which of the securities is the most significant in terms of market value and liquidity of the primary quotation of that security

[8] ISIN codes are issued at security level. That means a single ISIN code is issued for all listings of a share on all exchanges, provided these listings of the share are in its original form. If, however, a listing is in depository receipt or certificate form and is therefore a representation of the original share, the share qualifies for its own ISIN code issued by the numbering association of the country in which the depository receipt or certificate is issued. ISINID returns either P or S where P indicates that the equity record is the primary one (i.e., the domestic listing of the share or depository receipt or certificate), and where S indicates that the equity record is secondary (i.e., a foreign listing of a share or depository receipt or certificate)

[9] See Bartram et al. (2009) for a similar approach using Worldscope country identifiers. Out of the initial potential universe of constituents, Datastream carries 2136 missing values for the ICB classification, of which we were able to populate 58 from Bloomberg.

Applying filters i to viii above, the sample size decreases to 4092 stocks. At this point, mindful of the results of Ince and Porter (2006), we seek to clean the Datastream rough results, firstly by a review of the basic filters detailed above, secondly by introducing additional filters/criteria, both recursive and manual. This leads us to identify many more cases to exclude and a few to reinstate.

To review the basic filters, we do the following:

ix. When a matching ISIN or SEDOL is available, we first cross-check against Bloomberg the above criteria linked to security type, currency, ICB classification

x. We identify and exclude non-common stock constituents, mis-classified as common-stock, by searching for key words in their names[11]

The number of eligible stocks thus drops to 3127.

We have already encountered data errors at the stage of sample selection; there, correcting errors has been functional to establishing whether a stock possesses or not certain minimum requirements to join the final sample; with filters x to xvii, instead, we are already past that stage and we delve deeper into the time-series to deal with data errors at specific points in time for stocks that we have kept in the sample. We then apply additional filtering criteria:

xi. Without relying on Datastream date fields, we manually or recursively verify that returns/prices/shares information is present for at least 24 consecutive months, and that it is reliable. When missing, we try to fill in using Bloomberg data. We identify 24 months of data as a reasonable minimum, necessary to run robust and significant model estimation at monthly frequency. This filter eliminates a large subset of firms that make a brief apparition in the sample or that have bad or missing pricing and shares information

[10] Some authors like Brockman (2009) or Fu (2009) require 30 months, Arena (2008) requires 12 months, Bartram 25 weeks. We believe 24 months is reasonable choice in order to comfortably run estimation at monthly frequency

[11] Examples include keywords identifying preferred shares, unit or investment trusts, mutual funds. Collective investment funds are present massively, and have been identified looking for names of investment management firms

xii. To mitigate issues with quality of transaction data in Datastream, we investigate all gaps in the Volume series using Bloomberg, and fill in/correct them as much as possible

xiii. We exclude securities for which at least one of the following series is not available at all: returns, outstanding shares, volume or prices. Returns, prices and shares are the minimum set of data to run our analysis

xiv. We eliminate stocks that never reach the minimum market capitalization threshold (described below) of 144 Mln GBP. This level is identified as a threshold for inclusion in the sample, as discussed in greater details below (see footnote 17)

With these further screens, our final sample size drops to 1333.

The following tables summarize the impact of our sample selection and data quality filters on the final sample. Table I describes the impact of each group of filters, while Table V shows the detailed impact of some of the additional filters.

Table I - Impact of filters on sample constituents

	Included	Excluded
Full Universe	7968	
Basic filters (i to viii)	4092	3876
Basic filters reviewed (ix to x)	3127	4841
Additional filters (xi to xiv)	1333	6635

The table shows the incremental impact of data selection and data quality filters on restricting the number of eligible stocks included in the final sample from 7968 to 1333. The "excluded" columns shows incremental exclusions

Table II - Impact of additional filters on filtered sample

Additional filters	Excluded	In % of 3127
No Volume	332	11%
Below 144Mln	30	1%
No Shares	44	1%
Less 24 months	196	6%
Bad Shares data	1496	48%

The Table shows the separate impact of each "additional" filter on the sample already sieved by the "basic" filters (criteria i to x). The "excluded" column shows how many stocks are filtered out by each criteria alone, while the next column displays this figure as a percentage of the total number of stocks that comply with the basic data filters (3127). The "No volume" and "No Shares" criteria filter out stocks with missing volume or shares data. "Bad Shares data" captures cases with clearly wrong shares data. "Less than 24 months" eliminates stocks without returns/prices/shares information for at least 24 consecutive months. "Below 144 Mln" filters out stocks whose market capitalization never reaches 144 Mln GBP, a minimum size limit motivated in Section 5.3 hereafter

Table II shows very clearly the seriousness of the issues with Datastream data, and how they are concentrated on outstanding shares and volume information. It is important to stress that these figures are likely an understatement of the real underlying situation, as they refer to a sample that has already been substantially filtered, in terms of constituents, by more than 60% (from 7968 to 3127).

Throughout this process, Bloomberg data has been employed especially to verify and rectify unusual results of the sample selection screens described above, and of the additional data quality filters detailed already or hereafter. For instance, for all stocks with reported market capitalization above 5 Bln GBP[12] that failed one of the data quality or investability checks (described below in section 5.3), price, returns, volume and shares information has been manually checked against Bloomberg. This has helped identify and correct many data errors. Most common is the situation of incorrect share information carried by Datastream, which often fails to correctly account for the effect of stock splits, affecting in turn market cap and turnover calculations; numerous cases of incorrect volume, price and returns have also been found.

Data on Volume turned out to be particularly rich in gaps, and in order to reduce them we resort again to Bloomberg: we have first identified all stocks with market capitalization in

[12] At any point in time

excess of 200 Mln GBP[13] that have volume gaps in months with price information, then we have compared the volume series from the two providers, filling gaps and correcting outliers.

To gauge the extent of our corrective activity, we compare the raw data - from Datastream without any adjustment or correction - to our data after all the cross-checking and correcting effort. We do this exercise for the sample of stocks that pass all our filters detailed so far (1333) – thus for an already seriously filtered subset of the original -, and we show the results in Table III, which includes also data on price-to-book value as a control variable.

Table III - difference between raw and corrected database

Data type	Difference %
Return Index	12%
Price	27%
Shares	3%
Volume	32%
Price/Book	14%

The table quantifies the extent of corrections made to the raw Datastream data, for our final sample of stocks, as the percentage of all monthly datapoints that have been rectified across each variable

In Table III we show how much data differs between the raw and the corrected dataset, as a percentage of all potential data points (the product of the number of months times the number of stocks). The figures are quite substantial, especially for Volume and Price. However, this is an underestimation of the real difference, as it includes in the denominator also months with no data for a large number of stocks. A more precise estimation of the difference is shown in Figure 1, where we use as denominator only the number of stocks included in the sample according to the date fields, for each month[14].

[13] A very low threshold, resulting in an extremely thorough and lengthy – but by all means necessary in light of the results – reviewing process

[14] Figures can exceed 100%, when we fill-in data that was supposed to be missing according to Datastream date flags, but turns out to be available from Bloomberg

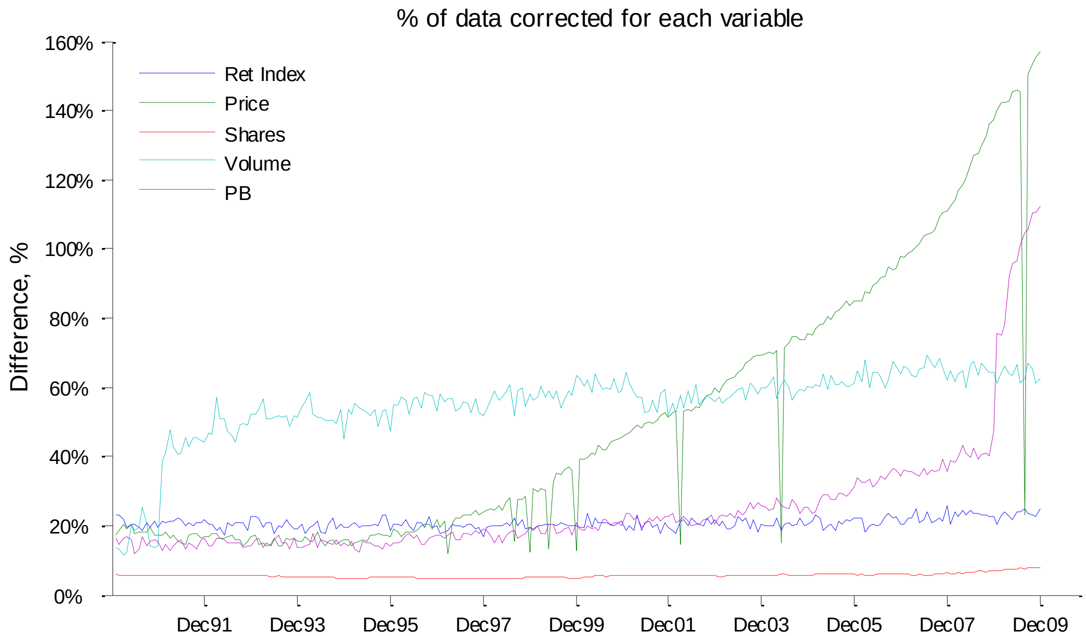

Figure 1 – Amount of corrections to raw data, as a % of active stocks in each month

The values shown in Figure 1 paint an even more serious picture; if the values for shares data range between 5% and 8%, the figures for Volume are stably around 60%. Although the extra large figures for Price can be downplayed a bit (they might partially be due to differences in the time or method of measurement), one key evidence is that the differences do not diminish over time towards the end of the sample: recent data from Datastream do not seem to warrant a lower level of attention than older ones.

Having determined our final sample constituents, we also apply the following filters on a month-by-month basis to take care of remaining data quality issues:

xv. If there is no change in the monthly return index, the stock is excluded from the working sample for that month

xvi. Subject to the previous requirement, we allow a maximum of 2 months of stable return index in the previous 24 month, provided the current month has a non-zero return[15]

xvii. We exclude monthly returns in excess of 200%

In order to strengthen the message on how serious the shortcomings of the Datastream samples are, we show in Table IV a selection of the constituents for which it has been necessary to correct, completely fill-in for or replace the Datastream values: they include some of the largest-capitalization stocks of the U.K. market.

Table IV - selected stocks with missing or incorrect data

NAME	Price	Return Index	Shares	Volume
BAT INDUSTRIES	1	1	1	1
BG GROUP	1	1		1
BP	1	1	2	
CADBURY		2		
COOKSON GROUP	1	1	1	1
CABLE & WIRELESS	1	1	1	1
INVENSYS			1	1
LLOYDS BANKING GROUP	1	1	2	1
ROYAL DUTCH SHELL B		1	2	1
SAINSBURY (J)		1		
VIRGIN MOBILE HOLDINGS	1	1	1	1
VODAFONE GROUP	1		2	1

1= missing or completely wrong in Datastream, sourced entirely from Bloomberg
2= Datastream values corrected by cross-check with Bloomberg or with information from company websites

What have we learnt from this preliminary filtering exercise? Two troubling lessons. First, within the Datastream sample, it is not possible to determine in a recursive, rule-based manner, which constituents must be retained and which should not. Static classification data is so incorrect that a review against alternative sources leads us to throw out circa 1000

[15] For all stocks with market capitalization of over 5 Bln GBP on the month when the above filters are triggered, we have manually verified data against Bloomberg overriding the (very few, 10) instances when the monthly returns happened to be actually be zero

constituents out of 4000. Second, historical data is not better: as shown in Table I, when we check the information carried on prices, shares, returns and volume, we have to throw out well more than half of an already restricted sample. The problem is not restricted to small stocks, but affects also large and mega-caps, especially as far as shares and volume data are concerned. Researchers should be aware of this, both when planning a cross-sectional analysis and when reading results drawn from samples that have not been treated accordingly.

5. Investability and liquidity filters

5.1 Method used to ensure investability and liquidity in previous literature

In addition to setting sound criteria for sample selection and data quality, we want to go further and ensure a rigorous minimum standard of investability and liquidity for all stocks in our filtered final sample, in order to make sure that statistical results have economic relevance for investors. We see investability as a broader concept than liquidity. We would characterize investable securities as those that can realistically and cost effectively be represented in institutional and pooled retail portfolios of reasonable size. We would therefore exclude from our sample stocks that fail one or more of a series of tests designed to identify potential hindrances to investment for such portfolios. We find that previous literature has usually approached this issue focusing on two data dimensions: relative market capitalization and non-trading days, with some authors considering also volume. A brief review follows.

Bali et al. (2005) test the robustness of Goyal and Santa Clara's findings with respect to several cross-sectional factors, primarily the impact of small/illiquid firms. In this context (rather than in a sample selection methodology context), after using a value-weighted methodology to question those results, they introduce screenings for size, liquidity and size, before running the one-month-ahead predictive regressions of the value-weighted portfolio returns on the recalculated volatility measures. As a size filter, they exclude, each month, all stocks with market capitalizations that would place them in the smallest NYSE size decile, choosing thus a mix of relative (in percentile terms) and absolute (because it uses NYSE stocks percentiles as a reference also for AMEX/NASDAQ stocks) size threshold. They give a good definition of liquidity as implying "the ability to trade large quantities quickly, at low cost, and without inducing a large change in the price level". They argue that trading volume is a natural measure of stock liquidity: as with size, they thus exclude all stocks that belong to the smallest NYSE volume decile on each month. In addition to the number of shares traded, following **Amihud (2002)**, they also measure stock illiquidity as the ratio of absolute stock return to its dollar volume (interestingly, introducing the price-size-liquidity screens, the predictive power of any volatility measure on the market returns disappears).

Guo and Savickas (2008), using Datastream data for non-U.S. G7 equity markets, employ some filters related to market capitalization data, although their exercise is centered on capturing errors in recorded data rather than assessing investability: they exclude stocks missing market capitalization data at the end of the previous quarter (their analysis runs at quarterly frequency), and they filter out excessive jumps (larger than 50% in absolute value) in daily market capitalization.

Ang et al. (2009), working on a Datastream universe of international stocks, exclude the 5% of firms with the lowest market capitalizations (apparently they define this filter in relation to the number of companies and not the aggregate market capitalization. Note also that they use this filter only for non-U.S. firms), obtaining an average number of firms equal to 1077 for the UK market. **Brockman et al. (2009)**, again using Datastream as data source for non-U.S. markets, exclude the firms making up the lowest 5% of the market cap, plus they go further requiring retained firms to have a minimum of 15 days a month with non-zero returns and positive volume. The 15 trading days per month requirement is also used by **Fu (2009)**. As we will discuss later, rules excluding the lowest (typically 5%) percentiles of cross sectional market capitalization or volume, might or might not be enough to remove stocks that typically are the constituents of “small-cap” indexes, depending on the structure of the market examined. At any rate, we think they might easily overlook other factors impacting “investability”. For instance, the portfolio side must be accounted for: absolute market capitalization will be important in defining investability if institutional investors including mutual funds were restricted (as they often are) from holding shares of a firm without a minimum liquidity. In this context, liquidity is usually defined as a minimum size and as the possibility to buy or sell a minimum quantity within a maximum time (provided by a decently deep book): thus absolute capitalization, absolute turnover and regular trading should be taken into account. Quoting MSCI’s methodology, “*The investable market segment includes all eligible securities with reasonable size, liquidity, and investability that can cost effectively be represented in institutional and pooled retail portfolios of reasonable size*”. In this respect, especially with reference to markets outside the U.S., and especially for non recent years, data quality is certainly an issue for researchers trying to gauge investability; for instance **Bartram et al. (2009)** are “forced” to use the frequency of non-trading as a proxy for market liquidity due to lack of reliable volume data: “*trading volume data at the firm level cannot be used because reliable trading volume data at the firm level are not available for a large percentage of our firm years”.* They emphasize this is a well-known shortcoming of the international returns data available from Datastream. Thus, they turn to return-based, alternative approach and filter out firms with more than 30% of zero-return weeks in a year. The situation might be more favourable for commonly used U.S. data, yet **Cremers (2007)** demonstrates the impact of filtering out firms with lacking turnover data on the U.S. CRSP sample size is not trivial: the reduction caused by the strictest filter, requiring no missing turnover data, is substantial, nearly halving sample size, and not decreasing in importance over time from 1967 to 2001.

5.2 Methods used to ensure investability and liquidity by leading index providers

For additional guidance, we thus turn to the methods used by the most important equity index providers, firms specialized in defining, calculating and providing indexes representing various segments of global and national stock markets: FTSE, MSCI, Dow Jones. A review of the liquidity rules used by the main index providers is doubtlessly important, not only because it covers the most effective methods developed as market standard practice, but also because these same methods are inextricably linked with the portfolio side of things, as most pooled investment vehicles are benchmarked and managed with reference to one index compiled accordingly. The index definition and filtering rules determine, de-facto, the investable universe for those investors, while passive management decisions, added to the strong recent growth in indexed products, imply that often the inclusion/exclusion from an index is a product as well as a cause of liquidity. Within the review, we pay special attention to the rules applicable to the U.K. market.

A first thing to note is the characterization of each market segment by capitalization percentiles. Although each national market has unique features, generally speaking the universe of stocks is split in four segments, large, medium, small and micro.

Generalizing again, the large cap segment makes up for 70-85% of the overall market capitalization, with mid-caps covering an additional 10-15%. This means that the “grey” area for small capitalization stocks generally starts from 85% of aggregate market cap, although recent (2009) FTSE UK index rules place it at about 95%.

It is worth stressing that the small cap segment is usually considered "investable", while the micro cap segment is not. The micro cap segment can be identified explicitly and/or implicitly: implicitly, through the use of market representation thresholds for the investable segments, usually set at 98-99% of the overall market capitalization; explicitly, as done by FTSE which puts in the micro (called "fledgling") segment stocks that fail some liquidity tests.

It is also important to note that those percentage numbers are a mix of deliberate choice and a result of the methodology used, which for the large and medium segments combines the use of a fixed number of constituents with explicit representation targets in terms of market capitalization. However, for the small cap segment, no market capitalization target is usually employed, determining a sort of "residual" definition: the small cap segment includes stocks that – while satisfying the liquidity requirements – are too small to be included in the previous aggregates, as opposed to the micro cap segment that must be more appropriately thought of as the aggregate of "illiquid" stocks. This should strengthen the concept that, as we move down the size scale, complementing market capitalization with other liquidity measures becomes more important.

What are then the liquidity filters used by the index providers?

FTSE (UK Indices): "Securities must be sufficiently liquid to be traded", in terms of price availability, size, liquidity. They define the large and mid cap segments as the largest 350 firms, and the small cap segment as the remaining companies that do not fail size or liquidity tests. They use a lower size limit of 0.2% (extended to as little as 0.05% for existing index members), defined as the ratio of the company full market capitalization to the full market capitalisation of the Small Cap index. Arguably, this relative size limit varies closely with the overall market capitalization and thus with prices, and seems quite loose: as they indicate that the sum of the large and mid cap segments account for about 96% of aggregate market capitalization, and that the small cap segment accounts for about 2% after applying the above tests, this suggest that the excluded lower end of the size spectrum of the potential universe is populated by a large number of micro firms (smallest than 0.2% of the lower 2%, that is, less than 0.4 bps of the aggregate market cap). As a liquidity test, they require index members to have a turnover of at least 0.035% of their shares in issue, based on their median daily trade per month (with zero trading days included), in ten of the previous twelve months. New issues are required to satisfy this condition on each month since their listing (annualizing the 0.035% median daily trade, we get a value close to 9%, that can be considered fairly loose), unless they are so large in terms of market capitalization (above 1% of the index) in which case they might be included anyway.

DOW JONES (Global Indexes series): they target to include all stocks with “readily available prices”. The market representation threshold, that is 98% for their more comprehensive “Global Total Market” series, becomes 95% for the “Global Indexes “series, which privileges investability. They allocate to the investable small cap segment from the 90th to the 95th cumulative percentile of total market capitalization; the lower 5% in terms of market capitalization is thus filtered out, but for the higher 95% no further size requirement is set. In terms of liquidity, stocks are screened for trading frequency, excluding stocks with more than 10 non-trading days during the quarter; this might be regarded as a quite loose constraint, as it would likely not capture situations with extreme swings in trading volumes.

MSCI (Global Investable Market Indices): they provide arguably the most coherent and convincing explanation for their methodology. The guiding principle, dictated by a strong emphasis on investability and replicability, is the definition of the “investable” market segment as including “all eligible securities with reasonable size, liquidity, and investability that can cost effectively be represented in institutional and pooled retail portfolios of reasonable size". A variety of factors are implicitly or explicitly taken into account to set forth the eligibility rules, including the number of companies, percentiles of market capitalization, absolute market capitalization level for the smallest company, the marginal contribution to the relevant index of the smallest company, the cumulative proportion of market capitalization covered, the liquidity and trading characteristics of companies, and an analysis of the average size of portfolio holdings of a variety of large, mid cap, and small cap investment managers.

For their family of Global indices, the "investability" tests are primarily a combination of size and liquidity requirements. The minimum size thresholds vary with prices and market developments, and are thus set as a function of the target representation coverage in terms of total market capitalization. The target coverage for the "Investable" universe is roughly 98% of the free-float adjusted total market capitalization (in line with DJ above); as of April 2010, this corresponds to a minimum size for inclusion in the investable index of 321 Mln USD of full market capitalization, with issues between 112 and 321 Mln USD considered non-investable. The Investable segment is split between the Large/Mid and the Small cap segments at the 85% of cumulative capitalization, a rule that yields a lower threshold for the Large/Mid segment at 1.76 Bln USD. The "Investable Small cap" space is thus between 300 Mln and 1.76 Bln USD. However, it must be noted that in order not to compromise the representation coverage, the thresholds are applied softly, within of range of 0.5 times to 1.15 times.

The liquidity requirements ensure a sound liquidity profile over both a long and short-term time-frame. To this end, they use the Annual Traded Value Ratio (ATVR), a measure of turnover that screens out extreme daily trading volumes, over both twelve months and three months, requiring at least a 20% ATVR over both timeframes. In addition they require eligible securities to trade at least on 90% of the trading days of a quarter, for 4 consecutive quarters. MSCI is the only provider to flag a potential liquidity issue caused by the "level" of a stock price: there may be liquidity issues for securities trading at a very high stock price. Hence, a limit of USD 10,000 is set and securities with stock prices above it fail the liquidity screening. However this rule is applied only for new issues from May 2008[16]. A minimum seasoning period of 4 month is also required, except in the case of IPOs worth at least 578 Mln for small caps or 3.16 Bln for mid/large caps. A seasoning period of 4 months of trading is required for IPOs, except for large ones, defined as those with market value of at least 1.8 times the relevant segment's minimum size. This would correspond to 1.8 times 321 USD Million = 578 USD Million for the Small cap segment, and to 1.8 times 1.76 USD Billion = 3.16 USD Billion for the large/mid cap segment.

5.3 Our filters for investability and liquidity

In the light of the above, the filters we employ are summarised as follows:

i. Minimum size limit: we use an absolute size threshold to ensure that stocks included in our sample would be tradable from the point of view of most institutional investors. To define it, we start from the limit of 321 Mln USD of total market capitalization used by the most recent MSCI methodology as of April 2010, and convert it to GBP using end of month exchange rates. Adjusting this limit back in time to account for exchange rates fluctuations, we obtain a range between roughly 160 and 230 Mln GBP. We thus use 160 Mln GBP as our minimum size threshold

ii. Once one issue reaches this threshold it is included in the working sample from that month onwards, unless its market capitalization falls more than 10% below the

[16] The MSCI U.S. domestic indices rules also quote potential problems caused by prices below 1 dollar, noting that this is already nearly always dealt with by exchanges through delisting. In a non-U.S. framework, this is an aspect to consider, for instance because extremely low prices could tend to generate disproportionately high percentage moves; we will thus use a lower limit at 10 GBp

threshold[17]. We introduce this -10% buffer to protect sample stability from market price fluctuations while maintaining liquidity[18]

In modelling the turnover constraints, we follow MSCI in taking into account both a shorter (monthly, quarterly) and a longer (yearly) horizon for measuring this dimension of liquidity. We want to allow some variability in monthly trading but also to ensure medium-long term liquidity. We thus impose a first 10% turnover constraint at a monthly level, but we require it to be satisfied over a rolling 12-month window; in setting this level of the monthly threshold, we use a value roughly equivalent, on an annualized basis, to the 0.035% daily turnover requirement currently used by FTSE. We then impose a second constraint (as noted earlier, the FTSE 0.035% daily requirement is quite loose) that uses MSCI threshold of 20% annualized quarterly turnover, over the past year. Our turnover constraints are thus:

iii. Minimum annualized monthly turnover of 10% in each of the past 12 months
iv. Minimum annualized quarterly turnover of 20% in each rolling quarter of the past year
v. Trading at least on 90% of the trading days in the previous 12 months, consistently with MSCI; evidence of a link between discontinuous trading and (il)liquidity can be found for instance in Bekaert et al. (2007)
vi. Issues trading below 10 GBp or above 10000 GBp will be excluded, to avoid both potential liquidity issues from very high or very low prices and abnormally large price jumps due to a very low, fractional price[19]

[17] This implies that stocks that never reach a market capitalization of 144 Mln GBP (90% of 160 Mlns) are never eligible for sample inclusion, hence the preliminary filter employed earlier (filter xiv in section 4.3)
[18] All index providers allow existing index members some degree of flexibility in satisfying chosen constraints, in order to safeguard sample stability
[19] Bali (2005), Arena (2008), Ruan (2010) use a 5$ minimum, while Ince and Porter (2006) use 1$ minimum. MSCI uses a minimum of 1$ and a maximum of 10,000$

All the above limits will work in combination. To give a rough idea of their impact in skimming the potential universe, the first three filters combined would narrow by 75% the number of eligible stocks from the FTSE Small Cap Index as of July 2010, bringing it down from 269 to 65.

5.4 Impact of filters on working sample

Figure 2 below shows the separate impact of each of the filters listed in Section 5.3 (plus the one based on basic data availability, i.e. the start and end dates for each stock's data series, and filters xv to xvii in Section 4.3) on the number of firms in the sample. Figure 3 instead shows their combined effect, and represents how many stocks pass all filtering criteria on each month.

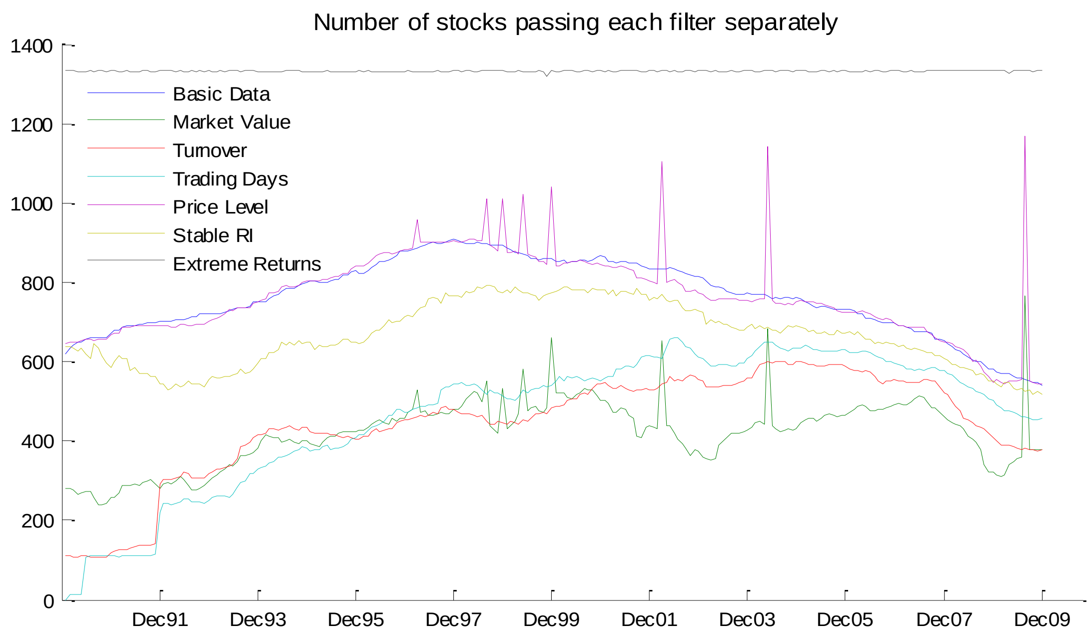

Figure 2 - Stand-alone impact of investability filters on working sample

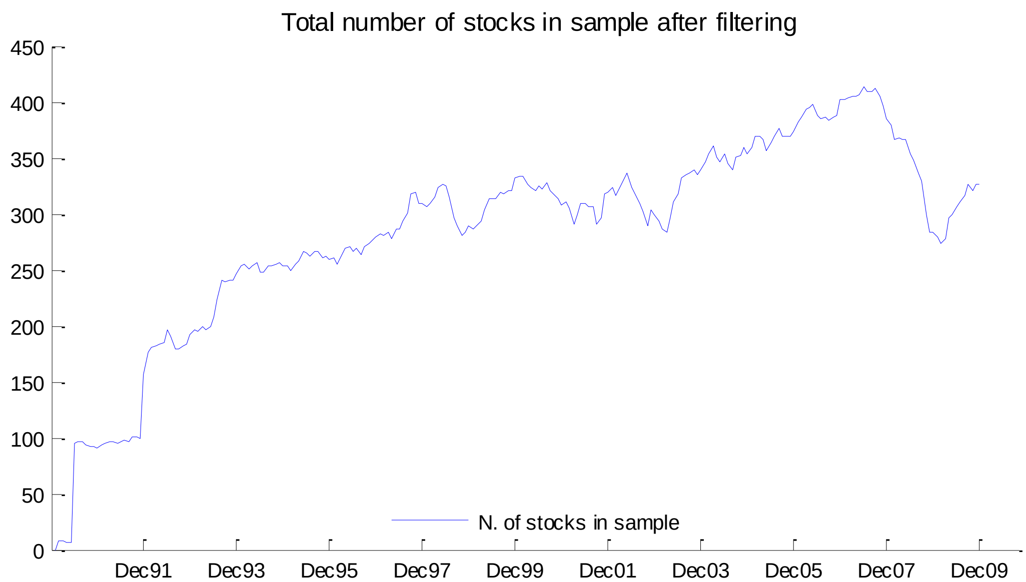

Figure 3 - Joint impact of investability filters on working sample

As can be easily spotted, there are obvious data-related issues with the Volume-based filters (based on turnover and trading days) in the initial part of the sample. Volume data carried by Datastream is particularly bad and scanty for the first 23 months, missing at least until the 23rd month for a significant number of stocks that would otherwise have been included in the aggregate filter, and giving rise to the clearly visible jumps in the number of eligible firms. Filling in with Bloomberg data does not help significantly for data going back to the early 1990's. In order to get around this issue, o reasonable approach is to lift volume-related filters for the first 23 months, recovering eligibility for a number of constituents. It is possible that doing this we could include firms whose low turnover in the initial months is not due to data issues; in this case we would just see them being thrown out of the sample in the 24th month, when volume filters come into effect, causing a sudden drop in the number of constituents. Thus we will impose the requirement that "recovered" stocks comply with both volume-related filters in the 24th month. The resulting final sample counts between 149 and 412 stocks per month, rising stably above 200 from the 41st month onwards. It is interesting to contrast our results with the number of stocks included in market index aggregates. We note that save for the period between October 2004 and June 2008, the final number of stocks satisfying our investability criteria lies below 350, the current target constituents number of the FTSE 350 Index, representing the Large/Medium Cap aggregate for the FTSE UK family of indexes. This is consistent with our screening criteria being more stringent, as it excludes away some stocks classified as investable, medium capitalization stocks under the FTSE methodology.

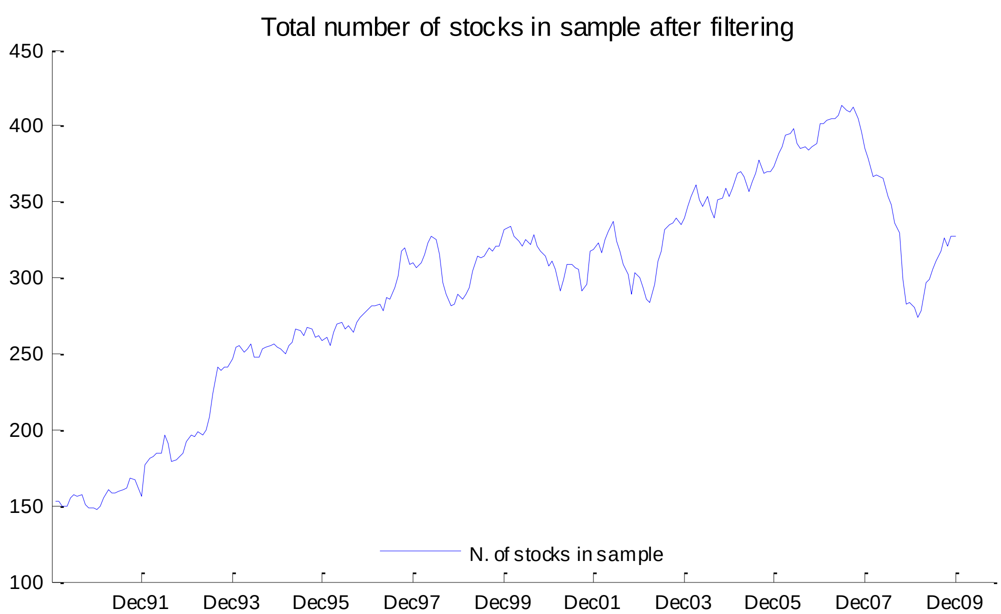

Figure 4 - Final working sample – Joint impact of investability filters

5.5 Sample descriptive statistics

In terms of market value, the minimum market capitalization of the firms included in our final sample remains very stable between 144 and 170 Mln GBP. The maximum size starts off around 30 Bln and gradually grows, reaching over 244 Bln on March 2000 - the tech bubble period saw a surge of corporate activity leading to large issuance and listings - then moderates towards 100 Bln. The mean market capitalization starts off at about 1.6 Bln and grows up to almost 5 Bln at the peak of the tech bubble, collapsing in the aftermath of the bubble burst and resuming an upward trend since 2003, getting just shy of 5 Bln again at the end of 2009.

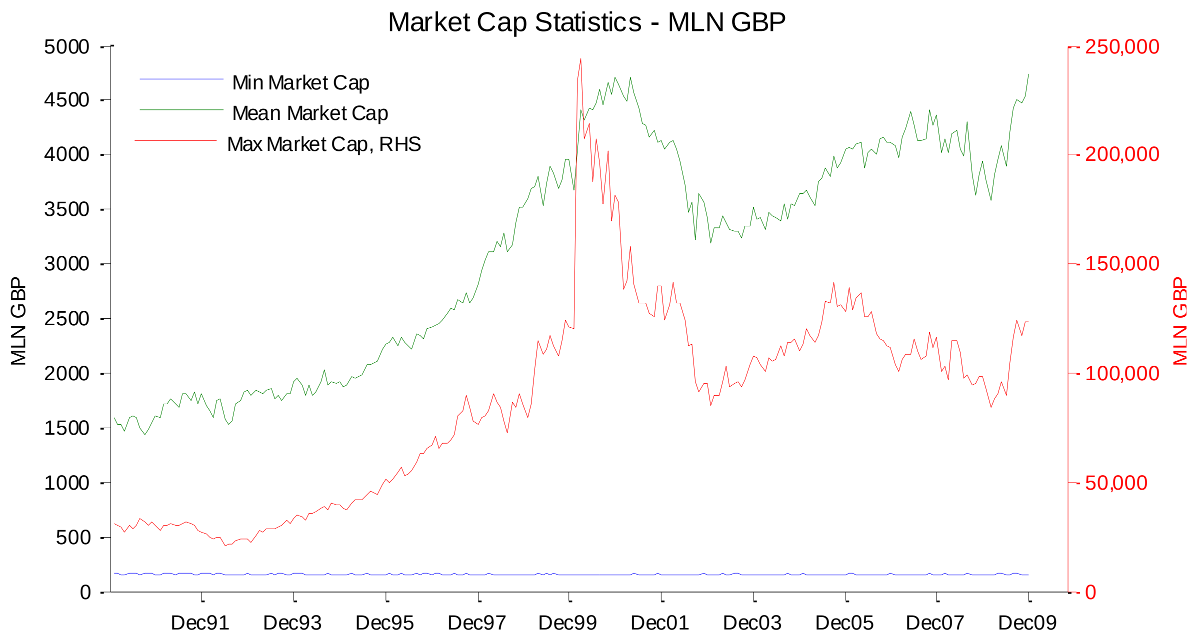

Figure 5 - Min, max and mean sample capitalization

The median size, however, is much lower and much more stationary, oscillating broadly between 600 and 800 Mln, although in recent years, from the start of the 2003 bull market, we note that it has trended up too, exceeding 1 Bln as of 31 December 2009, after a sharp and brief collapse due to the market rout following the financial crisis.

Naturally, mean capitalization tends to be affected by price behaviour more that the median; the stability of the median during the first three quarters of our investability-adjusted sample seems to indicate that the underlying market structure holds fairly constant in terms of size distribution[20].

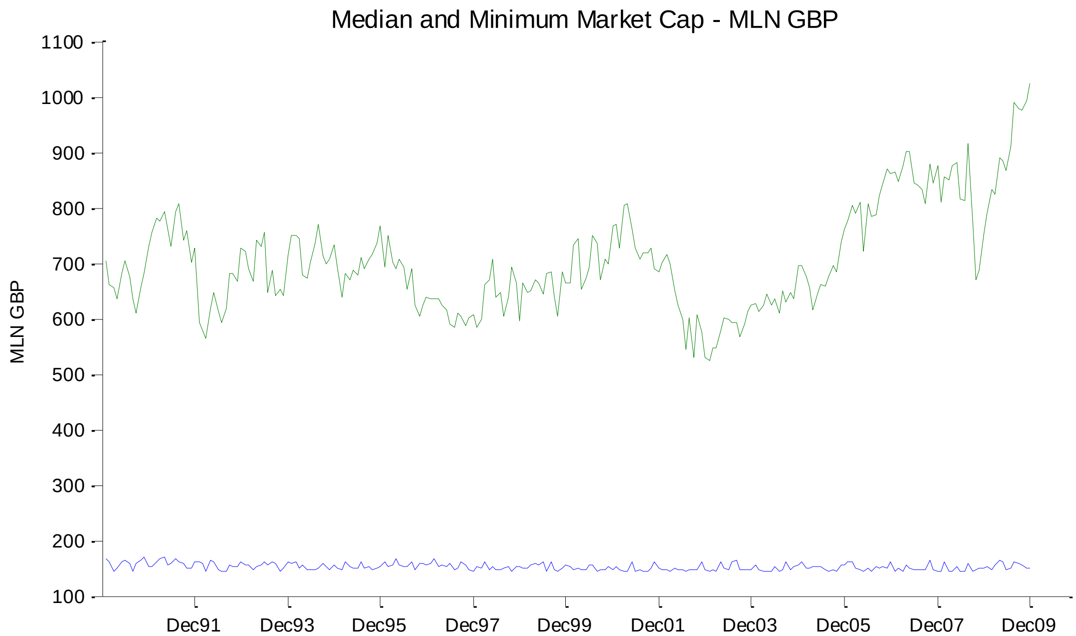

Figure 6 - Min and median sample capitalization

On the other hand, the upward drift in median size since 2003, during a period when the total number of firms rises only marginally, seems to point towards a tendency for consolidation among large player, determining a structural increase in the number of large (investable) firms relative to smaller (investable) ones. Median size almost doubles from its low of 520 Mln GBP on 31 Jan 2003, increasing over 20% from previous peaks, while the number of sample firms rises a “mere” 12% over the same period, remaining below previous peaks.

[20] Especially since the minimum size is floored at 144 MLN, while the maximum is not; the fact that the median size does not reflect the upward drift shown by the average size shows that the distribution of firms in the size percentiles is fairly stable

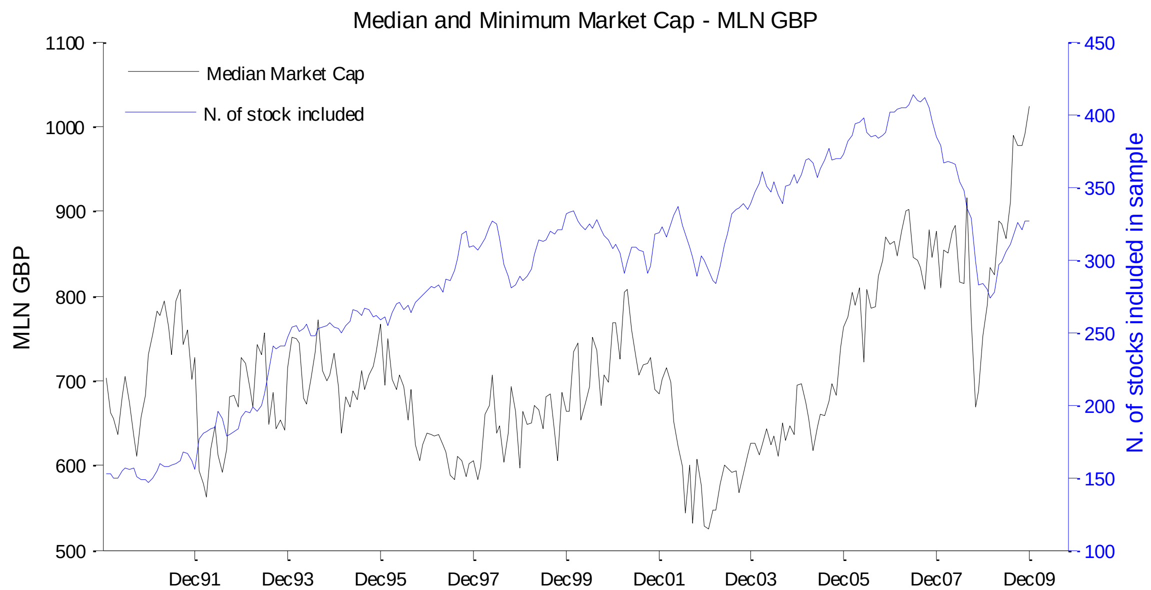

Figure 7 - Median capitalization and number of stocks

In Figure 8 we compare the aggregate market capitalization of our final filtered sample with the reported capitalization of the FTSE All Share Index. Our aggregate figure is necessarily smaller because of our stricter investability filters; yet the trend of the gap is worth a comment: it collapses over time, consistently with the joint effect of increasing data sample quality over time, and of increasing liquidity of the constituents stocks over time. The difference remains fairly large until 1998, and then quickly drops to reach levels close to 5% as early as mid 2002. The higher levels of the pre-2002 difference has two implications: a caveat that the early part of the sample, if not properly treated, suffers from serious data issues, and a confirmation that our filtering methodology is effective in mitigating them.

Further checks on alternative data sources for the aggregate market cap of the FTSE index (Bloomberg) indicate that, in addition, the reported number from Datastream for the early part of the sample might not be accurate, thus de-emphasizing the difference with our measure.

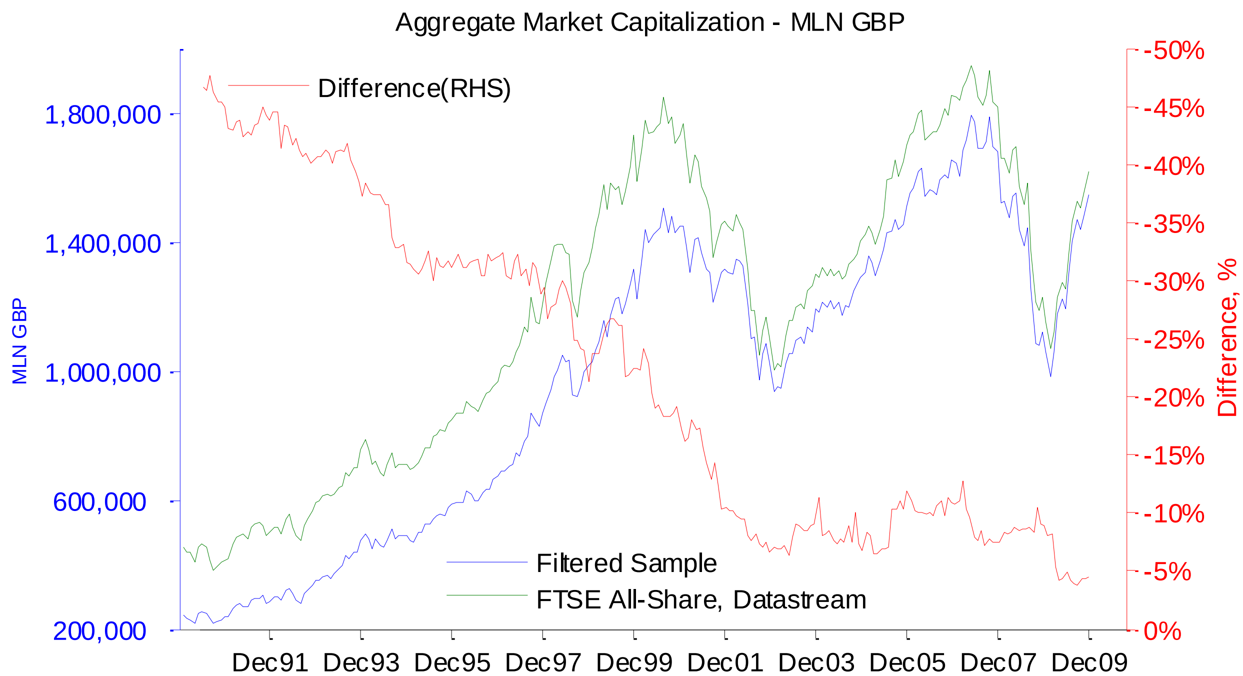

Figure 8 - Aggregate Market Value

Lastly, we employ the same criteria used by Ince and Porter (2006) to gauge the effectiveness of our filters in mitigating data quality issues, measuring the correlation of the returns of the relevant country market index with capitalization-weighted sample averages of returns built from our sample, using incremental applications of our filters. We show the results in Table V. The relevant country market index is the FTSE All-Share; at the stage when the sample has been already reduced to 3127 constituents applying filters data-quality filters i to x as described in section 4.3, the correlation between the FTSE returns and the market-cap weighted average returns of our sample is just 0.73[21]. Applying the additional data quality

[21] In addition, this value is biased upwards because at this stage we have already adjusted the reported start and end sample dates for a number of stocks

screens (filters xi to xvii) that bring the sample to the final number of 1333, the correlation becomes a more comforting 0.96. Finally, introducing the investability filters described in section 5.3, the correlation reaches 0.99.

Table V - Correlation of market cap-weighted sample return averages

	FT	MW_investable	MW_1333	MW_3127
FT	1.00			
MW_investable	**0.99**	1.00		
MW_1333	0.96	0.96	1.00	
MW_3127	0.73	0.69	0.77	1.00

Correlation between market-cap-weighted sample average returns and the returns of the FTSE All-Share Index (FT), based on monthly returns. “MW_3127” is the series of average market cap weighted returns computed after implementing the first set (i to x) of data quality filters, “MW_1333” uses all data-quality filters and “MW_investable” employs all data and investability filters, and has the highest correlation with the FTSE index returns, equal to 0.99

6. Conclusions

We add to the literature looking on cross-sectional equity analysis, by means of a comprehensive review of the Datastream U.K. sample, documenting serious issues of data quality, and dealing with them by means of a rigorous review and of an innovative methodology based on both data-quality and investability filters. We propose a novel approach to sample selection derived from best market practice in index construction and focused on investability.

Our review of U.K. listed stocks data sourced from Datastream shows it to be plagued by serious quality issues, in terms of both qualitative (classification) and quantitative (prices, returns, volume, shares) information, as highlighted in some previous work on international equity data. The problem is particularly severe for volume and shares information and pre-1999 data in general. We have found it possible to correct some of such instances by cross-checking qualitative information against an alternative source, mainly Bloomberg. However, we found that a large amount of incorrect data can be corrected only by manual, case-by-case verification, which we have undertaken. We have thus carefully reviewed methods used in previous literature to deal with issues of data quality and with the need to ensure appropriate liquidity for the chosen sample, but due to the fact that the vast majority of studies focuses on well-known U.S. samples, we have not found a great level of detail; the most careful approach to data quality issues is surely that of Ince and Porter (2006). We have thus applied filters to minimize remaining data-quality issues. Mindful of the importance of obtaining a dataset that could support economically relevant analysis and conclusions from the point of view of institutional market participants, we have devised investability criteria derived from the leading providers of market indexes. The innovation of this approach is that while previously documented sample selection criteria deal mainly with correction of errors and outliers, our "cleaned" sample retains only stocks that are likely to be "investable" and "tradable" for institutional portfolios of reasonable size, and we show this makes a significant difference restricting the eligible universe. We demonstrate that without our combined data-quality and investability filters the raw sample would suffer from serious data deficiency problems.

The result is a filtered sample that, on a value-weighted basis, shows a very high level of correlation with the published index for the U.K. market, improving significantly on the unfiltered sample, and should be able to support economically relevant cross-sectional statistical analysis.

References

Ang, Andrew, Robert J. Hodrick, Yuhang Xing, and Xiaoyan Zhang, 2009, High idiosyncratic volatility and low returns: International and further U.S. Evidence, *Journal of Financial Economics* 91, 1-23.

Ang, Andrew, Robert J. Hodrick, Xing Yuhang, and Zhang Xiaoyan, 2006, The cross-section of volatility and expected returns, *Journal of Finance* 61, 259-299.

Arena, Matteo P., K. S. Haggard, and Xuemin Sterling Yan, 2008, Price momentum and idiosyncratic volatility, *Financial Review* 43, 159-190.

Bali, Turan G., Nusret Cakici, Zhang Zhe, and Y. A. N. Xuemin, 2005, Does idiosyncratic risk really matter?, *Journal of Finance* 60, 905-929.

Bartram, Söhnke M., Gregory Brown and Rene M. Stulz, 2009, Why Do Foreign Firms Have Less Idiosyncratic Risk than U.S. Firms?, *Working Paper*.

Basu, Devraj, and Lionel Martellini, 2007, Total volatility and the cross section of expected stock returns, *EFMA Working Paper*, Draft.

Bekaert, Geert, Campbell R. Harvey and Christian Lundblad, 2007, Liquidity and Expected Returns: Lessons from Emerging Markets, *Review of Financial Studies* 20, 1783-1831.

Bekaert, Geert, Robert J. Hodrick, and Xiaoyan Zhang, 2008, Is there a trend in idiosyncratic volatility?, *Working Paper*.

Boyer, Brian H., Todd Mitton, and Keith Vorkink, 2008, Expected idiosyncratic skewness, *Working Paper*.

Brandt, Michael W., Alon Brav, John R. Graham and Alok Kumar, 2010, The Idiosyncratic Volatility Puzzle: Time Trend or Speculative Episodes?, *Review of Financial Studies* 23, 863-899.

Brockman, Paul, Maria Gabriela Schutte and Wayne Yu, 2009, Is Idiosyncratic Risk Priced? The International Evidence, *Working Paper*.

Campbell, John Y., Martin Lettau, Burton G. Malkiel, and Xu Yexiao, 2001, Have individual stocks become more volatile? An empirical exploration of idiosyncratic risk, *Journal of Finance* 56, 1-43.

Carhart, Mark M., 1997, On persistence in mutual fund performance, *Journal of Finance* 52, 57-82.

Chen, Joseph S., 2002, Intertemporal capm and the cross-section of stock returns, *SSRN eLibrary*.

Chua, Choong Tze, Jeremy Goh, and Zhe Zhang, 2006, Idiosyncratic volatility matters for the cross-section of returns - in more ways than one!, *Working Paper*.

Cremers, K. J. Martijn, and Jianping Mei, 2002, A new approach to the duo-factor-model of return and volume, New York University, *Working Paper*.

Cremers, K. J. Martijn, and Jianping Mei, 2007, Turning over turnover, *The Review of Financial Studies* 20, 1749.

Epstein, Larry G., and Martin Schneider, 2008, Ambiguity, information quality, and asset pricing, *Journal of Finance* 63, 197-228.

Fama, Eugene F. 1968a, Risk, Return and Equilibrium, *Center for Mathematical Studies in Business and Economics, University of Chicago*, Report No. 6831.

Fama, Eugene F., 1968b, Risk, Return, and Equilibrium: Some Clarifying Comment, *Journal of Finance* 52, 29-40.

Fama, Eugene F., and Kenneth R. French, 1992. The cross-section of expected stock returns. *Journal of Finance 48*, 427–465.

Fama, Eugene F., and Kenneth R. French, 2008, Dissecting anomalies, *Journal of Finance* 63, 1653-1678.

Frazzini, A., and L. Marsh, 2003, "Idiosyncratic Volatility in the US and UK Equity Markets," Unpublished Working Paper, Yale University.

French, K., G. Schwert, and R. Stambaugh, 1987, Expected stock returns and volatility, *Journal of Financial Economics*.

Fu, Fangjian, 2009, Idiosyncratic risk and the cross-section of expected stock returns, Journal of Financial Economics 91, 24-37.

Garcia, René, Daniel Mantilla-Garcia and Lionel Martellini, 2009, Idiosyncratic Risk and the Cross-Section of Realized Returns: Reconciling the Aggregate Returns, *Working Paper*.

Guo, Hui and Robert Savickas, 2008, Average Idiosyncratic Volatility in G7 Countries, *Review of Financial Studies* 21, 1259-1296.

Guo, Hui, and Robert Savickas, 2004, Idiosyncratic volatility, stock market volatility, and expected stock returns, *Working Paper*.

Hamao, Yasushi, Jianping Mei, and Yexiao Xu, 2003, Idiosyncratic risk and the creative destruction in Japan, *NBER Working Paper*.

He, Zhongzhi, Sahn-Wook Huh and Bong-Soo Lee, 2008, Is Dynamic Factors and Asset Pricing, *Working Paper*.

Ince, Ozgur S., and Burt R. Porter, 2006, Individual equity return data from Thomson Datastream: handle with care!, *Journal of Financial Research* 29, 1475-6803.

Jacobs, Kris, and Kevin Q. Wang, 2004, Idiosyncratic consumption risk and the cross section of asset returns, *Journal of Finance* 59, 2211-2252.

Jegadeesh, Narasimhan, and Sheridan Titman, 2001, Profitability of momentum strategies: An evaluation of alternative explanations, *Journal of Finance* 56, 699-720.

Jensen, Michael C., Fischer Black and Myron Scholes, 1972, The Capital Asset Pricing Model: Some Empirical Tests, *Studies in the theory of capital markets* 81, 79-121.

Lehmann, Bruce N., 1986, Residual risk revisited, *NBER Working Paper*.

Lehmann, Bruce N., 1992, Empirical testing of asset pricing models, *NBER Working Paper*.

Li, Xiafei, Joëlle Miffre, Chris Brooks, and Niall O'Sullivan, 2008, Momentum profits and time-varying unsystematic risk, *Journal of Banking & Finance* 32, 541-558.

Malkiel, B. G., and Yexiao Xu, 2002, Idiosyncratic risk and security returns, *Working Paper*.

Malkiel, Burton G., and Yexiao Xu, 1997, Risk and return revisited, *Journal of Portfolio Management* 23, 9.

Merton, Robert C., 1987, A simple model of capital market equilibrium with incomplete information, *Journal of Finance* 42, 483-510.

Ruan, Tony, Quian Sun an Yexiao Xu, 2010, When Does Idiosyncratic Risk Really Matter?, *Working Paper*.

Spiegel, Matthew, and Xiatong Wang, 2005, Cross-sectional variation in stock returns: Liquidity and idiosyncratic risk, Yale School of Management, *Working Paper*.

Wei, Steven X., and Chu Zhang, 2005, Idiosyncratic risk does not matter: A re-examination of the relationship between average returns and average volatilities, *Journal of Banking & Finance* 29, 603–621

Xu, Yexiao, and Burton G. Malkiel, 2003, Investigating the behavior of idiosyncratic volatility, *Journal of Business* 76, 613-644.