EVALUATING THE PERFORMANCE OF AGGREGATE SURVEY-BASED INDICATORS IN TIMES OF CRISIS.

Andreas Belegratis
Statistician & Data Scientist

Confederation of British Industry, CBI

Final Report submitted for ECFIN-BCS Workshop

October 21, 2016

Abstract

In a data driven economy, qualitative business survey data are of great importance. Business and Consumer Surveys have proven to contain a vast amount of vital information and, when designed and harvested correctly, can give an indication of what is happening ahead of corresponding official data.

This report discusses "*How Survey-Based Indicators perform in time of severe economic crises and what are the implications*". The statistical techniques that will be used include the survey balance, the probability and the regression methods which serve as the basis for quantifying survey-based expectations, along with perceptions of the past and provide relatively "early" economic indicators. This is then followed by an application on the Confederation of British Industry's Industrial Trends Survey with the focus on using these survey data to provide early economic indicators that track the UK's manufacturing output both in normal times and in times of crisis. To test the performance of these indicators during these different periods, many *out of sample* forecasting experiments were conducted on the financial crisis period, from 2008Q2 to 2009Q2. We find that the Industrial Trends Survey data significantly outperform benchmark ARMA models when forecasting UK's manufacturing output both in normal times and in crisis. We also examined the impact of changes in sample size and in answering practices of the Industrial Trends Survey during normal and crisis periods, and the experiment confirmed the robustness of survey data. The above results highlight the importance and robustness of ITS survey data that provide two early indicators for output, one as a nowcast and one as a forecast as well as the utility of the quantified series to be used as efficient predictors in other more complex macroeconomic models to increase their forecasting accuracy.

KEYWORDS: Quantification, Aggregate Survey Data, Survey-Based Indicators, Quantified Expectations, Performance Evaluation, Crisis, Forecasting Experiment.

Address for correspondence: CBI, Cannon Place, 78 Cannon Street, London, EC4N 6HN, U.K. Tel: +44 (0)746 915 5274, +44 (0)783 726 3482, e-mail:Anna.Leach@cbi.org.uk, Andreas.Belegratis@cbi.org.uk.

Contents

1. Introduction

The subject of measuring survey-based macroeconomic expectations is still relevant and a highly important subject amongst scholars, economists and survey practitioners - especially since the global financial crisis. Economists around the globe are fighting a constant battle to produce timely and precise forecasts to reduce the uncertainty about the future of the overall economy. Extreme abnormal events such as the 2008 financial crisis by their very nature are difficult to anticipate by using the predictions of business agents at an aggregate level.

Although the financial crisis originated in the US housing market, it soon spread to UK mortgage lenders and led to the first run on a British bank (Northern Rock) in over 150 years. Official statistics for Gross Domestic Product, as provided by the UK Office of National Statistics, show that the subsequent recession in the UK lasted for the five quarters 2008 Q2 to 2009 Q2 until growth was observed once again in 2009 Q3.

The focus of this report is to provide a theoretical and a practical guide on how to construct early aggregate economic indicators and assess their forecasting abilities in times of crisis. To do that, an experiment will be conducted whereby we will go back in time to 24 April 2008 to when the Confederation of British Industry quarterly Industrial Trends Survey (ITS) was published, and try to forecast UK manufacturing output using a combination of the available official statistics and survey data at the time. The latest official statistics on manufacturing for 2008 Q1 were not published until mid-May 2008[1]. As a result of the publication lag, one can use the retrospective views of ITS respondents and give an early indicator for what has already occurred during 2008 Q1 about a month before the official data for 2008Q1 are published. Also one can use the prospective (expectations) views of the respondents to provide an early forecast for what will happen in 2008 Q2.

In order to be able to provide these early indicators and help economists understand what has happened to the economy up to that time and make timely forecasts, business and consumer tendency surveys are used as a tool to gather data by asking questions of individual respondents such as consumers, firms, governments, etc. regarding their past and/or future views on various topics, such as their volume of output, prices, employment, exports and many other characteristics associated with various key economic indicators such as inflation, output growth and GDP.

Several countries in the EU have long standing so called "tendency surveys". These kinds of surveys usually ask respondents to report *"Down"* or *"Fall"*, *"Same"* or remain *"Stable"*

[1] For this report it is assumed that at quarter t the only official data available are the data published for the previous quarter $t-1$.

and *"Up"* or *"Rise"* in an ordered fashion, concerning their individual expectations on the future movement on prices, volume of output, employment rate, etc. In the United Kingdom, the longest standing UK economic survey is the Industrial Trends Survey (ITS) managed by the Confederation of British Industry (CBI), which began the survey in 1958 and asks manufacturing firms questions on past and future views on various micro and macroeconomic topics.

The structure of this paper is as follows: in section 2, starts by outlining and motivating the theory of the most known *quantification* methods the *balance statistic*, the *probability approach* and the *regression approach* that are proposed in the literature since the early 1950's; in section 3, the CBI's Industrial Trends Survey is introduced, the dataset is described and the *in-sample* and *out of sample* sets are defined as well as results from the *in-sample* analysis are summarised. These two sets will be the basis for the *real time crisis experiment*. In section 4, the *real-time crisis experiment* is described and the *out of sample* analysis results are outlined, as well as a series of experiments are conducted to test the forecasting ability of survey-based expectations and the potential impact of changes in sample size during normal and crisis periods. The conclusion involves a general discussion about the utility and performance of survey data and their quantified proxies served as Survey-Based Economic Indicators in normal times and in times of crisis.

2. Quantification Methods

The three fundamental approaches that are most commonly used in literature, to quantify survey-based expectations using *aggregate* survey information are the *balance*, the *probability* and the *regression* methods.

Consider a survey that asks at time t, N_t respondents on their opinions about the actual (population) value of an economic variable X_t^*, and let the actual change of that underlying variable from $t-1$ to t be ${}_{t-1}x_t^*$ or simply x_t^*.

Furthermore, let the survey at time t ask agents two questions regarding the movement of an economic variable X_t^*, one is retrospective and concerns x_t^* and the other is prospective and concerns the future trend x_{t+1}^*. The answers available for each respondent $i = 1, .., N_t$

are usually three e.g. expect X_t^* or x_{t+1}^* to go "*Down*", remain the "*Same*" or go "*Up*". There is also a "*Not Applicable*" option which is supressed for the purpose of the analysis[2].

Furthermore, let the population change of X_t^* from $t-1$ to t denoted as x_t^* to be a weighted average of all agents in the industry (population) meaning $x_t^* = \sum_{j=1}^{N^*} w_{jt} x_{jt}$ where w_{jt} is the weight of each agent (e.g. firm) in that particular industry with population size N^* (at time t) and x_{jt} denotes each agent's individual change e.g. in output from $t-1$ to t.

In order to provide an early indicator for the overall change x_t^* or value X_t^* a survey is used. Hence, when the survey asks a *sample* of N_t firms (agents) from the industry (population) of N^* firms, to express their qualitative views about the future change of their e.g. prices, output etc. from (t) to $(t+1)$ denoted as x_{it+1}, then one can estimate the actual future change x_{t+1}^* of X_t^* from a weighted sample average $x_{t+1} = \sum_{i=1}^{N_t} w_{it} x_{it+1}$. The component x_{it+1} is not yet observed because is the future change e.g. of the i^{th} firm's e.g. prices, output over the next period. Although it is not observed, surveys collect qualitative data by asking agents to provide an expectation about the movement of x_{it+1}. Hence, one can use these survey expectations data gathered from each i^{th} agent, in order to estimate x_{it+1} as x_{it+1}^e then estimate x_{t+1} as x_{t+1}^e which by itself is an indicator for the (population) change x_{t+1}^*. [3] The problem that arises is how to get from agents' qualitative future views ("*Rise*", "*Fall*", "*Same*") to quantitative measures x_{it+1}^e about their own actual future change $x_{it+1}$and thus obtain an average measure x_{t+1}^e about x_{t+1} which can be though as an early economic indicator about the future change of X_t^* (x_{t+1}^*).

There are various ways outlined in the literature that one can use when facing the problem of quantifying qualitative survey-based expectations which are known to be used as Economic Indicators for the future of the underlying variable[4] x_{t+1}^* or X_{t+1}^*. Despite the fact that exact individual expectations ${}_t x_{i,t+1}^e \coloneqq E[\ {}_t x_{i,t+1} | \Omega_{it}]$ [5] cannot be directly computed, because of the qualitative nature of the survey data, an approximation for those i individual expectations

---

[2] The percentage of firms reporting "Not applicable" is considerably small, less than 1% and will be ignored by allocating the percentage equally to the other three answers see Berk (1999). Also it is assumed that answers "Up" with "*Rise*", "Down" with "*Fall*" and "Same" with "*Stay Stable*" are respectively equivalent.

[3] X_{t+1}^* and x_{t+1}^* refer to the actual value and percentage change in population (whole market). Whereas, the X_{t+1} and x_{t+1} are the corresponding sample estimations. For example X_{it+1} is the future value of output that a respondent firm will produce, over the next period $(t+1)$ and also the firm's output growth from (t) to $(t+1)$ is denoted as x_{it+1}. The objective is how to estimate X_{t+1}, x_{t+1} (quantitative) from survey data (qualitative) and thus have an estimation for X_{t+1}^* and x_{t+1}^*.

[4] Depending on the question asked in the survey the underlying variable could be the change $(x^* \coloneqq \Delta X^*)$ or the value (X). Furthermore the change from $(t-1)$ to (t) could be e.g. $x_t \coloneqq \Delta X_t = X_t - X_{t-1}$ or $\frac{X_t - X_{t-1}}{X_{t-1}}$ or $\log\left(\frac{X_t}{X_{t-1}}\right)$. In this report the underlying variable in question is the percentage change x, $x_t \coloneqq \frac{X_t - X_{t-1}}{X_{t-1}}$ where t symbolizes each quarter.

[5] ${}_t x_{i,t+1}^e$ is the expectation of the i^{th} respondent agent standing at time t, for the movement of the underlying quantitative variable x (percentage change) regarding the next period $t+1$ given all the information available up to time t.

can be used. How close the approximation will be to the actual change highly depends on the quality of survey data and the model assumptions that have to be made in order to obtain a measure of expectation for the future change of the underlying variable[6].

Usually, survey data are published in an aggregate form basically aggregate percentages of "Ups", "Downs" or ("Ups" − "Downs"). In order to convert the qualitative expectations to quantitative measures three basic approaches will be used that deal with the aggregate form of the survey data.

1. Balance method, Anderson (1951)
2. Probability method, Theil (1952) and Carlson & Parkin (1975)
3. Regression method, Anderson (1952) and Pesaran (1984)

2.1 Balance method

The first approach of quantification is the so-called *balance statistic* B_t which was coined by Anderson (1951) and is the difference in percentages of survey participants who responded positive to those who responded negative. In this case the difference between the percentage of firms at time t who expect "*Rise* " and those who expect "*Fall*" concerning the movement of the underlying variable x over the next period.

Let the aggregate form F_t^e, S_t^e and R_t^e represent the total percentage of firms expecting x to "*Fall*", stay *"Stable"* or *"Rise"* over the next period with codes 1, 2, 3 respectively. Notice that the survey is conducted at quarter $t-1$ and the percentages concern the quarter t. Below the definitions for F_t^e, S_t^e, R_t^e are outlined.

- ${}_{t-1}F_t^e$ or F_t^e: Percentage of firms expecting their output to "Fall" from $t-1$ to t. Note that the same principles apply for ${}_tF_{t+1}^e$ or F_{t+1}^e only the time index changes from t to $t+1$ and so forth for every t.
- ${}_{t-1}S_t^e$ or S_t^e:Percentage of firms expecting their output to stay "Stable" from $t-1$ to t.
- ${}_{t-1}R_t^e$ or R_t^e: Percentage of firms expecting their output to "Rise" from $t-1$ to t.
- $F_t^e = \frac{\sum_{i=1}^{N_{t-1}} \mathbf{I}\{r_{it}^e=1\}}{N_{t-1}},\ S_t^e = \frac{\sum_{i=1}^{N_{t-1}} \mathbf{I}\{r_{it}^e=2\}}{N_{t-1}},\ R_t^e = \frac{\sum_{i=1}^{N_{t-1}} \mathbf{I}\{r_{it}^e=3\}}{N_{t-1}},$

  $F_t^e + S_t^e + R_t^e + NA_{t-1} = 1$

[6] In the scope of simplifying the notation, always consider x_t to be the actual total growth from the previous period ${}_{t-1}x_t$ until it is stated otherwise. This means when writing x_{t+1} is the same as ${}_tx_{t+1}$ and so forth, same applies to the individual growths x_{it}. The right hand side always denotes the time the underlying variable is attributed to.

Where r_{it}^{e} is the coded answer {1, 2, 3} for agent i and NA_{t-1} denote the "Not Applicable" answer to expectations question in the survey $t-1$ a and $\mathbf{I} = \begin{cases} 1, & if\ r_{it}^{e} = j\ \ and\ j = 1,2,3 \\ 0, & otherwise \end{cases}$ is the index function that basically counts each code in the survey $t-1$ in order to get the aggregate percentages.

The balance statistic was originally formed for the trichotomous case where only three responses were available but it can be generalized to the pentachotomous case (or more) simple by using the appropriate weights.

$$B_t = -1 * SD_t - 0.5 * D_t + 0 * \ S_t + 0.5 * I_t + 1 * SI_t$$

Where $SD_t < LD_t < S_t < LI_t < \ SI_t$ are ordered and represent "strong decrease", "light decrease", "stable", "light increase", "strong increase". The above balance statistic is by itself an average quantified measure for the actual (future) change x_t . Although has to be scaled in order be an economic indicator and track the actual change of X e.g. output or inflation rate or an index. The choice of the scale will also affect the accuracy of the Balance Statistic to track the official data. Then Balance statistic could be considered as the mean of a common discrete distribution for all agents were answers are located in points $(-\theta, 0, \theta)$.

(0) $x_t^{BAL} := B_t = \theta(R_t^e - F_t^e)$ and $\hat{\theta} = {\sum x_t}/{\sum(R_t^e - F_t^e)}$

The scaling parameter θ is estimated as $\hat{\theta},$ by a-priori imposing that the quantified expectation series x_t^{BAL} is an unbiased estimator for the mean of the underlying variable x_t over the whole sample period which can be shown to be rather restrictive since unbiasedness is a necessary condition for expectations to be rational and would be better if it was not imposed a-priori (see Batchelor & Orr 1988). Although many different scaled parameters have been proposed in the literature one can bypass the scaling problem by standardising both the x_t^{BAL} and x_t and focus on their correlation. The balance statistic could be considered as the mean of a discrete aggregate probability distribution with a dispersion measure (variance) $DIS_t = \hat{\theta}((R_t^e + F_t^e) - (R_t^e - F_t^e)^2)$, where answers are located to the points of -1 for a "*Rise*" 0 for a "*Fall*" and 1 to "*Stay the same*". The fact that the distribution of forecasting series come from a discrete distribution (see Batchelor 1986) became the motivation for Theil (1952) to find a more flexible approach which lead to the next method of discussion which is the probability method.

2.2 Probability method

The probability method although was first seen in Theil (1952), it was popularised later by Carlson and Parkin (1975) in their attempt to quantify inflation expectations using the CBI survey data.

The method is based around the assumption that each i^{th} individual's response, forming an expectation at time t $_{t}x^{e}_{i,t+1}$ about the future movement of $x_{i,t+1}$ (e.g. i^{th} firm's individual output) comes from a subjective conditional probability density function $f_i(x_{it+1} \,|\Omega_{it})$ with mean $_{t}x^{e}_{i,t+1}$ and a dispersion measure $_{t}\sigma^{e}_{i,t+1}$ where Ω_{it} is all the information available to agent i up to time t. Additional assumptions should be made for the first and second moments $E[f_i(x_{t+1} \,|\Omega_{it})] < \infty,$ $E[f_i(x_{t+1} \,|\Omega_{it})]^2 < \infty$ to be finite in order for the distribution to have a mean and a dispersion measure such as variance.

Because each i^{th} respondent is expected to report at $(t+1)$ the mean of the above probability distribution it holds that each response $j = 1, 2, 3$ is constructed as follows, agent i reports :

- "*Fall*" , if their expectation is below $_{t}x^{e}_{i,t+1} \leq -\beta_{it}$
- "*Stable*" , if their expectation is between $-\beta_{it} <$ $_{t}x^{e}_{t+1} < \alpha_{it}$
- "*Rise* ", if their expectation is above $_{t}x^{e}_{i,t+1} \geq \alpha_{it}$

Where the threshold parameters $a_{it}, \beta_{it} > 0$ form an interval $[-\beta_{it},\ \alpha_{it}]$ which in bibliography is called the "indifference interval" or "indifference limen" or "just noticeable difference" as shown in Figure 2.1 for the Carlson & Parkin (1975) case.

The interpretation of that interval comes from the assumption that when firms are forming expectations in their mind two significant thresholds exist, one positive α_{it} and one negative $-\beta_{it}$ where they will report they expect a *"Rise"* if $_{t}x^{e}_{i,t+1} \geq \alpha_{it}$ and a "Fall" if $_{t}x^{e}_{i,t+1} \geq \alpha_{it}$ if they consider for example that, according to their personal views of the market and their own finances their output is unlikely to increase or decrease from (t) to $(t$+1$)$ then it holds $_{t}x^{e}_{i,t+1} \in (-\beta_{it},\ \alpha_{it})$ and they should report remain "*Stable*".

Carlson and Parkin (1975) furthermore assume that the firms are *independent* with each other and *identically distributed* then the individual subjective conditional probability distribution $f_i(x_{t+1} \,|\Omega_{it})$ can be aggregated to form the joint conditional probability distribution $f(x_{t+1} \,|\Omega_t) = \prod_{i=1}^{N(t)} f_i(x_{t+1} \,|\Omega_{it})$ where $\Omega_t = \bigcup_{i=1}^{N(t)} \Omega_{it}$ is the available information to all firms up to time t.

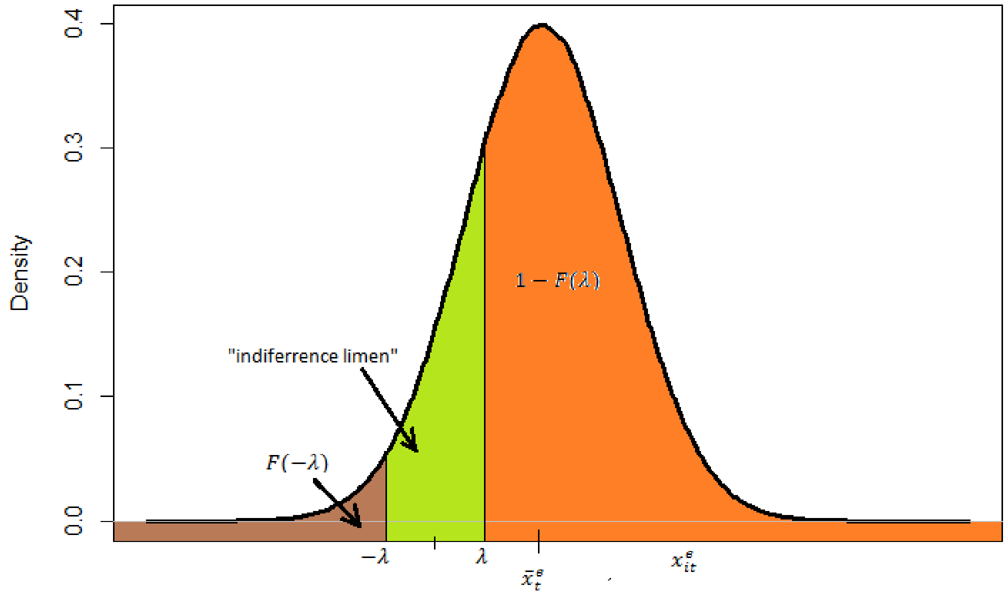
Now consider the individual expectations series $x_{i,t}^{e}$ $\forall i$ as independent identically distributed draws from the joint probability distribution with mean $x_{t}^{e} \coloneqq E\left[\ x_{i,t}^{e} \middle| \ \Omega_{t}\right]$ and variance $\sigma^{2}{}_{t}^{e} \coloneqq V\left[\ x_{i,t}^{e} \middle| \ \Omega_{t}\right]$ or standard deviation σ_{t+1}^{e}.

Furthermore, Carlson and Parkin (1975) assume the distribution of x_t, F to be the normal distribution density function. They also assume that the threshold parameters are symmetric, constant and time invariant $a_{it} = \beta_{it} = \lambda > 0$ all across agents. Thus, the percentage of sample of firms reporting *"Rise"* and the percentage of those reporting *"Fall"* from $t-1$ to t, converges in probability to true population values as $N_t \rightarrow \infty$ (large enough). This means that the probability to observe a future *"Rise"* and *"Fall"* is approximated as:[7]

(1) [8] $R_{t}^{e} \rightarrow P(x_{it}^{e} \geq \lambda) = 1 - F(\lambda)$

(2) $F_{t}^{e} \rightarrow P(x_{it}^{e} \leq -\lambda) = F(-\lambda)$

2.1 The distribution of mean expectations: The Carlson & Parkin (1975) method



The percentage $S_{t}^{e} = 1 - R_{t}^{e} - F_{t}^{e}$ of individuals reporting no change is used to define the indifference limen $-\beta_t < \ x_{t}^{e} < \alpha_t$ or as in the special case of Carlson and Parkin (1975) $-\lambda < \ x_{t}^{e} < \lambda$. By standardisation in (1) and (2) we get

---

[7] The probability to observe no change on x_t is $S_t \rightarrow P(-\lambda < {}_{t-1}x_{t}^{e} < \lambda)$ where ${}_{t-1}x_{t}^{e}$ should lie inside the symmetric and constant indifference interval $(-\lambda, \lambda)$.

[8] $(F_{t}^{e}, S_{t}^{e}, R_{t}^{e})$ correspond to the question on expectations and (F_t, S_t, R_t) correspond to the question on past realizations of x_t

$$(3)\ 1 - R_t^e = \Phi\left(\frac{\lambda - x_t^e}{\sigma_t}\right)$$

$$(4)\ F_t^e = \Phi\left(\frac{-\lambda - x_t^e}{\sigma_t}\right)$$

Where Φ is the cumulative probability function of the standard normal distribution $N(0,1)$. Thus by inverting (3),(4) and taking the quantile function which is the inverse function of the cumulative probability distribution function or $Q(x) = \Phi^{-1}(x)$ an explicit solution is derived.

$$(5)\ \sigma_t = \frac{2\lambda}{Q(1-R_t^e) - Q(F_t^e)}$$

$$(6)\quad x_t^e = \lambda\left(\frac{Q(1-R_t^e) + Q(F_t^e)}{Q(1-R_t^e) - Q(F_t^e)}\right)$$

Where Q is the quantile function of the standard normal distribution. Also, note in this specific case where the indifference interval is considered symmetric and time invariant λ merely becomes a scaling parameter for the x_t^e to track the actual expected realisations of x_t.

Notice that (5) and (6) are two equations with three unknowns x_t^e, σ_t and λ. Carlson and Parkin estimate λ by imposing a-priori unbiasedness over the whole *in-sample* period meaning that $E[\lambda\, x_t^e] = E[x_t]$ thus,

$$(7)\ \hat{\lambda} = \frac{\sum_{t=1}^{T} x_t}{\sum_{t=1}^{T}\left(\frac{Q(1-R_t^e) + Q(F_t^e)}{Q(F_t^e) - Q(1-R_t^e)}\right)}$$

This assumption of a-priori unbiasedness is criticised from many authors. Mostly because unbiasedness is a necessary condition for the expectations to be rational and it should not be imposed a-priori when quantifying expectations. Ideally the unbiasedness property of x_t^e should be tested after the quantification procedure and not be initially imposed. Anyhow the survey based indicator formed at $(t-1)$ for (t) given by the Carlson & Parkin (CP) probability method can be summarised as:

$$(8)\ x_t^{CP} = \hat{\lambda}\,\frac{f^e + r^e}{f^e - r^e} \qquad \text{where } f^e = Q(F_t^e) \text{ and } r^e = Q(1 - R_t^e).$$

Since Carlson & Parkin (1975) various extensions have been proposed in the literature such as different probability distributions for x_t e.g. central t that has heavier tails and give x_t a higher probability to take more extreme values. This feature is useful especially in crisis because of the large negative values of x_t. Other examples include uniform, non-central t, logistic, χ^2 and many others see Batchelor (1981), Mitchell (2002), Nielsen (2003) et.al. Also, Carlson & Parkin (1975) themselves consider the central t distribution as an alternative. Studies show (including ours) that the results of using other distribution do not differ significantly enough even when normal distribution assumption is clearly violated e.g. 2008-2009 financial crisis. Although there are not much evidence to suggest taking a normal distribution is not appropriate there are have been many studies that criticize the CP assumption of the indifference interval $[-\lambda, \lambda]$ to be time invariant across surveys and symmetric between respondents and many extensions have been proposed in the literature for a quick overview and shortcomings of the Carlson & Parkin method see Nardo (2003). One can relax the above assumption of the interval to an asymmetric one $[-b, a]$ a time varying one $[-\lambda_t, \lambda_t]$ or both $[-b_t, a_t]$ which the general case. Mitchell (2002) proposes an even more generalized version for the indifference interval $[-b_{it}, a_{it}]$ to also vary across each respondent, but that will require the analysis of panel data or disaggregate which is not on the scope of this report. Another problem is that this method is developed for the trichotomous case when there are only three possible answers available to respondents and two threshold parameters but in a more general scenario where the pentachotomous (5 answers) or polychotomous (n-odd answers) then the assumptions of symmetry and time invariance for the indifference limen are strongly violated.

As an alternative to bypass the generalisation problem and the a-priori unbiasedness Pesaran developed the regression method.

2.3 Regression Approach

This method has its roots in Anderson (1952), when he wanted to find another justification for the balance statistic and was further developed and applied by Pesaran (1984) who used it as an alternative measure to the probability method in order to forecast inflation rates with CBI ITS data. The basic idea is to use the relationship between respondents retrospective views $(R_t\ S_t\ F_t)$ on past realisations and official data x_t to serve as a 'yardstick' for quantifying the respondents expectations (prospective views) (F_t^e, S_t^e, R_t^e).

Let us again consider the underlying variable x_t as a weighted average of respondents perceptions[9]

(9) $x_t = \sum_{i=1}^{N_t} w_{it} * x_{it}$

where w_{it} is the weight attributed to the i^{th} respondent. Assuming that (9) holds for the sample of respondents participating in the survey t. Then, by categorising the respondents on who

[9] x_{it} could be either retrospective or prospective time notation changes.

reported a *"Rise"* as (+) and *"Fall"* as (−) on the movements of the underlying variable x_t, (9) can be rewritten as follows:

(10) $x_t = \sum_{i=1}^{N_t} w_{it}^{+} * x_{it}^{+} + \sum_{i=1}^{N_t} w_{it}^{-} * x_{it}^{-}$

Anderson (1952) assumes the relationship between agents reporting an increase $(+)$ or decrease $(-)$ with the underlying variable moves around a constant. This means that each agent's magnitude of reporting $+$ and $-$ on average remains constant and time invariant across all agents.

(11) $x_{it}^{+} = a + v_{it}^{-}$

(12) $x_{it}^{-} = -\beta + v_{it}^{+}$

Where v_i^-, v_i^+ the independent and identically distributed zero mean error terms $(E[v_i^-] = E[v_i^+] = 0)$ with constant (and relatively small) standard deviations σ^+, σ^- across all firms $(\forall i)$ and over the whole sample period $(\forall t)$. Then, by substituting (11) and (12) on (10) and by using as weights $\sum_{i=1}^{N_t} w_{it}^{+} = R_t$ and $\sum_{i=1}^{N_t} w_{it}^{-} = F_t$ the percentages of firms that reported "*"Rise"* " and "*Fall*" for the previous period about x_t one has following linear regression:

(13) $x_t = aR_t - \beta F_t + v_t$ *Anderson's Model*

Where $a, \beta > 0$ and $v_t = v_t^- + v_t^+$ is the disturbance term. Pesaran (1984) while attempting to forecast inflation criticised the appropriateness of Anderson's model by arguing that one should expect an asymmetric relationship between the rate of change individual agents prices and the average inflation rate. Pesaran modified (11) to allow for an asymmetric relationship.

(14) $x_{it}^{+} = a + \lambda_1 * x_t + v_{it}^{-}$

(15) $x_{it}^{-} = -\beta + v_{it}^{+}$

(16) $a, \beta > 0$ and $0 \leq \lambda_1 \leq 1$

By substituting (14)(15) to (10), (13) becomes (17):

(17) $x_t = \frac{a\, R_t - \beta F_t}{1-\lambda_1\, R_t} + v_t$ *Pesaran's Model*

$$v_t = \frac{(\sum_{i=1}^{N_t} w_{it}^{+} v_{it}^{+} + \sum_{i=1}^{N_t} w_{it}^{-} v_{it}^{-})}{1-\lambda_1 R_t}$$

Equation (17) is a nonlinear regression possibly autocorrelated and heteroskedastic through R_t. Smith & McAleer (1990) extended Pesaran's model (17) to also allow the asymmetric relationship both ways thus (14), (15), (16) become:

$$x_{it}{}^{+} = \quad a + \lambda_1 * x_t + v_{it}^{-}\ , \qquad x_{it}{}^{-} = -\beta + \lambda_2 * x_t + v_{it}^{+}$$
$$0 \le \lambda_1 \le 1, \qquad\qquad 0 \le \lambda_2 \le 1 \qquad\qquad a, \beta > 0$$

(18) $$x_t = \frac{a\, R_t - \beta\, F_t}{1 - \lambda_1 R_t - \lambda_2\, F_t} + v_t$$ *Smith & McAleer's Model*

$$v_t \;=\; \frac{(\sum_{i=1}^{N_t} w_{it}^{+} v_{it}^{+} + \sum_{i=1}^{N_t} w_{it}^{-} v_{it}^{-})}{1 - \lambda_1 R_t - \lambda_2\, F_t}$$

Pesaran assumes that either (13) or (17) (18) holds not only for realisations R_t, F_t but also for expectations R_t^e, F_t^e. Then, if these regressions do not show any autocorrelation in their error term, then one can obtain an average measure of quantified expectations by simply taking expectations on the (13) (17) (18) because the error term has mean zero.

(19) $$x_t{}^{AND} = \hat{a}\, R_t^e - \hat{\beta} F_t^e$$

(20) $$x_t{}^{PES} \;=\; \frac{\hat{a}\, R_t^e - \hat{\beta}\, F_t^e}{1 - \widehat{\lambda_1}\, R_t^e}$$

(21) $$x_t{}^{SMAC} = \frac{\hat{a}\, R_t^e - \hat{\beta}\, F_t^e}{1 - \widehat{\lambda_1}\, R_t^e - \widehat{\lambda_2} F_t^e}$$

The estimations for $\hat{a}, \hat{\beta}\,, \hat{\lambda}_1, \hat{\lambda}_2$ are the OLS estimations of the linear regression (13) (or nonlinear (17) or (18) depending on which model fits the data best) of respondents retrospective views and past realisations of official data x_t. Pesaran (1984) also proposed an AR(*p*) structure on the error term v_t to account for the potential autocorrelation and Smith & McAleer (1995) consider a MA(*q*) as an alternative. By allowing $v_t \sim AR(1)$ (20) becomes

(22) $$x_t{}^{PES} \;=\; \frac{\hat{a}\, R_t^e - \hat{\beta}\, F_t^e}{1 - \widehat{\lambda_1}\, R_t^e} + \hat{\phi} v_{t-1}$$

$\hat{\phi}$ is the parameter estimated from AR(1) on the residuals given from (17). To conclude, a possible setback of the regression method is the autocorrelation in the error term v_t in (13)(17)(18). Furthermore, the numerical estimation of the nonlinear regressions (17)(18) with autoregressive errors might not converge especially for a high order ARMA autocorrelation structure, as well as the interpretation of the parameters is not clear anymore. Pesaran (1987) described (17) as not a traditional regression but is simply used to identify the relationship between two different pieces of information. In this section the most known quantification methods for aggregate survey data were discussed along with proposed extensions.

To summarise, the quantification methods can be applied to either the prospective (expectations) data to obtain a quantified measure of *forecast* for x_{t+1} or on the retrospective data to obtain a *nowcast* for x_t. One can also do a combination of both as seen above in the regression method. After each quantification procedure and depending on the type of data (or combination) used one should obtain an early indicator either a forecast for the next period or a nowcast for the past period which is not yet officially published. One also should remember that depending on the quantification method the "correct" scaling should be applied in order to "track" the official data or an index based on the official data. One example in the regression method (13)(17)(18) is that an intercept is added to ensure the quantified series are unbiased estimates and are tracking x_t. For the Balance and Probability method when the scaling is wrong, in order to obtain an indicator one could try re-scaling the quantified series by regressing it on official data series x_t.

Once the quantified expectation measures are obtained from each method discussed in Section 2. One could be tempted to investigate if the expectation series exhibit any rationality. An expectational variable such as ${x_t}^{BAL}, {x_t}^{CP}, {x_t}^{AND} {x_t}^{PES} {x_t}^{SMAC}$ is said to be (strictly) rational if the following four conditions hold unbiasedness, lack of serial correlation, efficiency and orthogonality. Plainly speaking this means that the agents' expectations should be unbiased estimations of official figures, while efficiently refers on agents ability of using all the available information when forming expectations. Roughly, for expectations to be weakly rational should be no evidence of autocorrelation on the disturbance term, the expectations have to be unbiased estimators of the underlying variable and agents do not form expectations just by using past values of the underlying. This was observed before in the regressions were it was a necessary condition in order to go from (13)(17)(18) to (19)(20)(21).

The error contained in the survey data after the quantification procedure can be broken down into three components:

- $x_t^{quant\_nowcast} = x_t - e_{1t}$
- $x_t^{quant\_exp} = x_t^{actual\_exp} - e_{2t}$
- $x_t^{actual\_exp} = x_t - e_{3t}$

To test the R.E.H one needs to examine the behavior of the quantified expectation forecast against the underlying realizations meaning $u_t \coloneqq x_t - x_t^{quant\_exp}$ or better

- $u_t = e_{2t} + e_{3t}$

This means that the behavior of u_t is influenced by the size and systematic nature of the error coming from the conversion method. Thus if the conversion error is significant enough one could get false positive results.

To test the unbiasedness hypothesis for the quantified expectations (any method) consider for example the following regression:

(23) $x_t = c + b\, x_t^{CP} + e$

Then parameters in regression (23) ($x_t^{quant\_exp} = x_t^{CP} from\,(8)$) should not be found to statistically differ from $c = 0$, $b = 1$. Be careful because it is not individual t-tests, it is necessary for the $c = 0$ and $b = 1$ at the same time. The t-test evaluates the hypothesis of $H_0 : c = 0\; given\; b = 1$ is in the model (23) we need $c = 0$ and $b = 1$ to be tested simultaneously. Hence, F-test is used for a joint hypothesis testing. The problem is the asymptotic results from F-test hold only when all conditions of linear regression are fully satisfied which is not usually the case. Actually to test both unbiasedness and autocorrelation at the same time in the regression (24) $[c = 0\,, b = 1, \varphi = 0]$ have to hold simultaneously.

(24) $x_t = c + b\, x_t^{CP} + \varphi e_{t-1}$, $e_t \sim AR(1)$

To test for $\varphi = 0$ one can use the so called "runs test" or Kolmogorov-Smirnov test or a similar test that can distinguish the systematic nature of the error term.

Instead of x_t^{CP} one can use any other indicator obtain from other quantification methods such as x_t^{BAL}, x_t^{PES}etc. in order to test the R.E.H.

One can also test the (weak form) efficiency condition by regressing the forecast error $u_t = x_t - x_t^{CP}$ against past values of x_t.

(25) $u_t = c + b\, x_{t-1} + e$

The above equation if $b = 0$ indicates the forecasting error is orthogonal to information regarding past values of x_t thus x_{t-1} does not help to improve the forecast which essentially means that firms have used "all the available information" efficiently. As you may suspect x_{t-1} is not the only (common) relative available information to all firms at time t. Hence, one in order to test the strong condition of R.E.H. needs a wider set of relative and commonly available information to all firms in that market.

(26) $u_t = c + b\, \Omega_{t-1} + e$

For further details see Batchelor (1982) (1986)(1988), Lee (1994), Pesando (1975) and Pesaran (1989) and for a quick summary of results by many researchers see Nardo (2003). This concludes our part of the theory and now seems appropriate to continue with an application on the CBI's Industrial Trends Survey.

3. Industrial Trends Survey : Application on UK's manufacturing

3.1 Outline of the ITS Survey

The Industrial Trends Survey (ITS), by the Confederation of British Industry, is the longest running survey on the UK manufacturing which began in 1958 and continues to be an accurate and timely bellwether for UK manufacturing sector and the wider economy. The ITS asks manufacturing firms key questions on the *past* (nowcast) and *future* (forecast) regarding the movement on domestic and export orders, capacity, output, employment, investment, competitiveness, optimism, training and innovation. The firms have three responses available: *"Up"* which indicates a *"Rise", "Same"* or *"Stable"* and *"Down"* which indicates a *"Fall"* on the underlying variable x_t. There is also a "*Not Applicable"* option which will be supress for the analysis.

The questions chosen for this analysis from the quarterly Industrial Trends Survey are:

Question (**8a**)*: "Excluding seasonal variations, what has been the trend over the* ***past*** *three months, with regard to the* ***volume of output****?"*

1. *"Down"*
2. *"Same"*
3. *"Up"*

Question (**8b**)*: "Excluding seasonal variations, what are the* ***expected*** *trends over the* ***next*** *three months, with regard to the* ***volume of output****?"*

1. *"Down"*
2. *"Same"*
3. *"Up"*

(**8a**) refers to respondents' retrospective views and (**8b**) to their prospective views. The prospective views will be quantified into quantitative forecasts (economic indicators) using the methods discussed in section 2. The retrospective views will serve as a 'yardstick' in the regression method to quantify the survey expectations (prospective views) and obtain an average measure for the next three months $(t + 1)$ volume of output (x_t). Another use for the retrospective data will be mentioned later in section 4 when the *out of sample* experiment is conducted.

3.2 Dataset description and matching

The ‘*whole’ sample* of the dataset used for this analysis consists of 74 quarters from 1991Q1 to 2009Q2 which is the end of the financial crisis. Over the *whole sample* period the ITS has an average sample size of 985 respondents (firms) in the manufacturing sector. The goal of this report is to forecast the financial crisis on manufacturing sector from start to finish using the ITS data. Thus, the *whole sample* is split between two sub samples, the *in sample* period 1991Q1:2008Q1 with average response sample of 1017 firms and the *out of sample* period 2008Q2:2009Q2 with an average of only 551. The *in sample* refers to the data available to someone in the moment the quarterly ITS for April 2008 is published. The *out of sample* period includes the last 5 quarters refers to the data that one would attempt to forecast.

In order to compare what manufacturing firms report and expect and what actually happened, matching with the official data is crucial. The matching between the survey and official data requires connecting the ITS with the MPI. The MPI is a monthly and quarterly survey managed by the United Kingdom’s Office of National Statistics (ONS) which has a sample of around 9000 firms and collects firm level quantitative data on turnover from all sectors and uses to estimate the national GDP. ONS publishes times series of official data by industries. The index used to track the output growth is the Index of Production which includes Manufacturing we need the latter is because ITS asks only manufacturing firms.

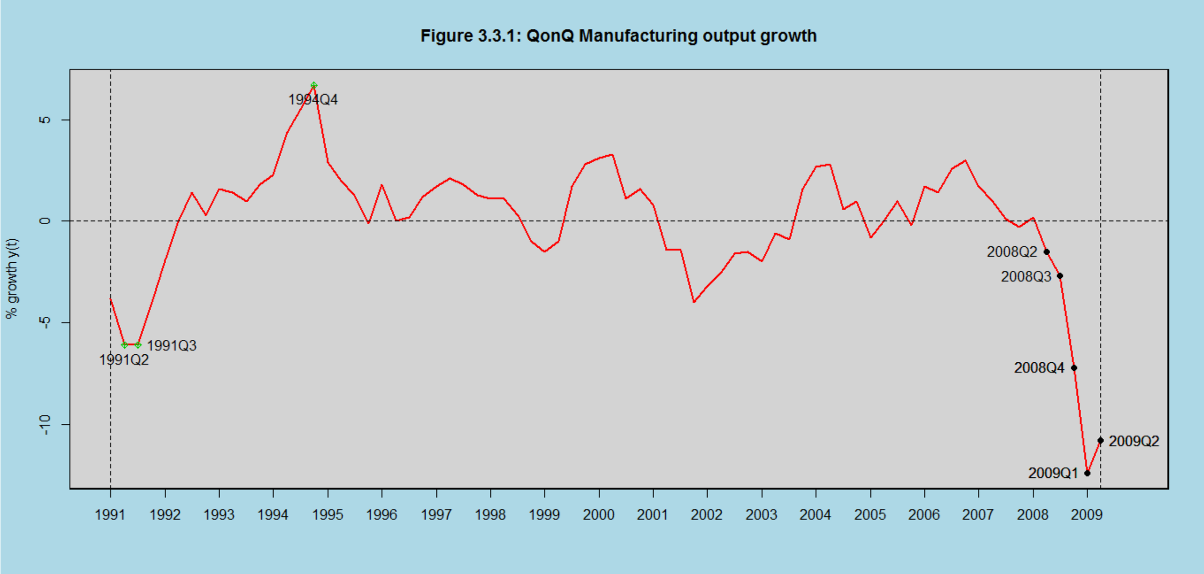
ITS retrospective, prospective views correspond to QonQ growth on an annual rate. Thus, to align and match survey data and quantitative data the QonQ Manufacturing output growth on an annual rate time series was chosen from ONS. Now meaningful comparisons can be made because the manufacturing firms’ nowcasts as well as the expectations when quantified will be used to track the QonQ manufacturing output growth time series provided by the ONS. An example of the dataset can be shown in the Appendix Table B. In order to match survey and official data as effectively as possible one has to also consider the time lag of publication between the two surveys. To highlight a simple example regarding the publication lag the ITS published in late April 2008 (basically start of 2008Q2) asks firms to report what has been the trend of the output on the previous three months essentially 2008Q1 and what is expected to happen on the next three months essentially[10] the 2008Q2. The ONS on the other hand waits for all the 2008Q1 to end then sends the questionnaires and publishes the results for 2008Q1

[10] ITS sends questionnaires about 1 week before the end of the month and collects them back about 1.5 weeks of the next month and publishes the results in end of that current month. ITS published in 24th of April 2008 but the questionnaires collected around the 10-14nth of April. This means that the most part of ITS April 2008 retrospective question (8a) corresponds to the 2008Q1 official data which would have been published by the ONS on May 2008 one month after. On the prospective question (8b) largely corresponds to a forecast on 2008Q2 even though some firms responded in early 2008Q2 which is the start of April. Theoretically there is a gap because some firms respond before the quarter ends and some after but this is assumed not to be significant.

in mid May 2008. This basically shows the significance of the ITS[11], is that it provides a more timely indicator for manufacturing output considerably ahead of the ONS data for the same period.

3.3 Descriptive statistics

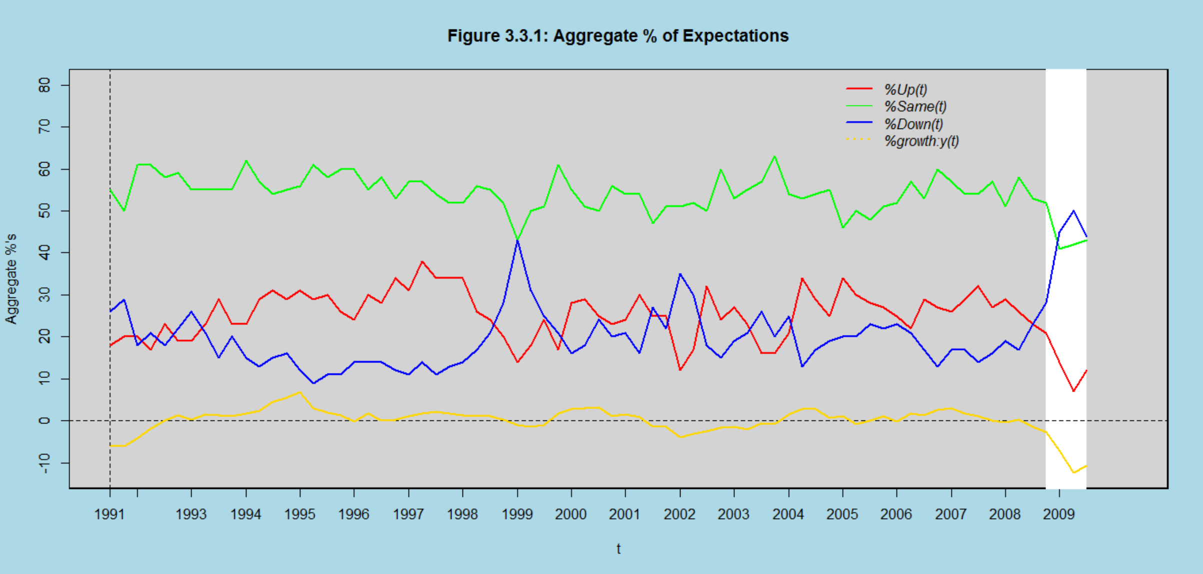
The matching of the survey data with official data is completed. From Figure 3.3.1 The QonQ Manufacturing % output growth is relatively smooth over the whole sample period except form the period before 1993 and the after the 2007 which correspond to crisis periods. The *out of sample* data for the output are shown as the black dots. The green dots symbolise the outliers in the *in-sample* period. UK was also in a recession during the early 1990's which lasted until 1991Q3.



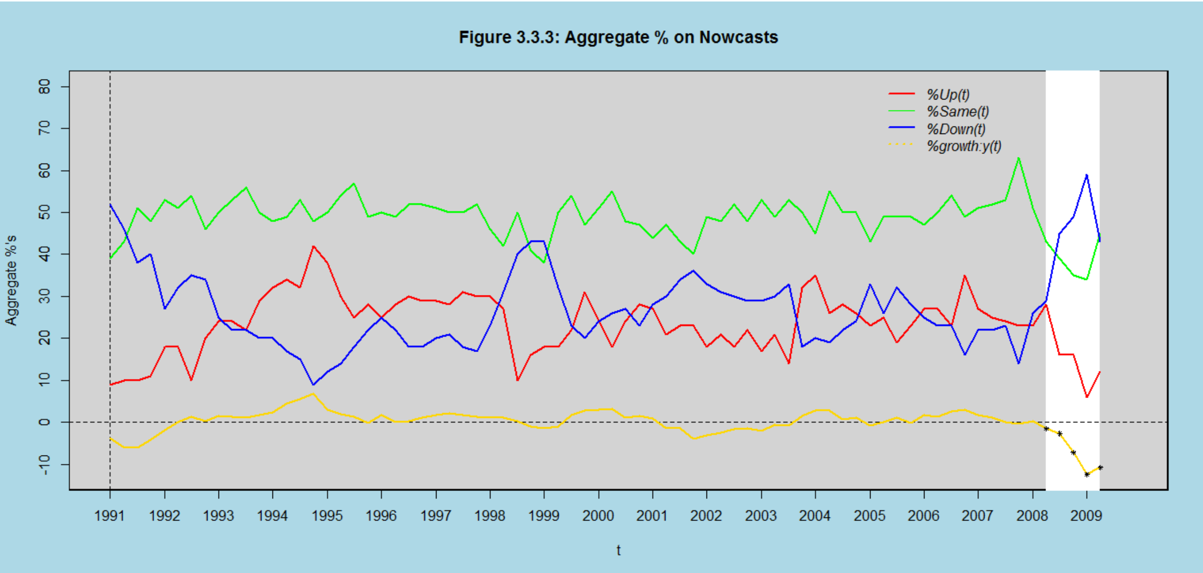
Comparing this with the ITS Figures 3.3.1 and 3.3.2 show the percentages of firms who answered questions 8a and 8b which are oddly high during the whole sample period. As far as the expectation series is concerned in 3.3.1., firms do not seem to expect the output to decline that low in the quarters 1991Q2 and 1991Q1 as the percentages firms reporting "*Up*" and "*Down*" are pretty close to each other. In mid 1998-1999 firms expect a considerable decline in growth (blue line larger than red) but it does not actually occur (dotted yellow line are the official statistics). The white part of the graph is the financial crisis period. The % of firms expecting output will remain the *"same"* is has declined considerably since 1998Q2 meaning that more firms felt the pressure. In the 2009Q1, the deep recession is observed it is also picked

[11] Note that once CBI publishes ITS on April 2008 corresponding to 2008Q1, one could provide an early estimate for what ONS will publish in May 2008 for 2008Q1 by using the firms' retrospective views from ITS April 2008 and also by using the prospective views one could provide a forecast for the 2008Q2 official data that ONS will publish on August 2008.

up by the ITS where % of firms in 2008Q4 expecting "Up" on "2009Q1" hits a record low of just 5% and the % of firms expecting the output to go "Down" hits a record high almost 50%.



Let's turn now to what firms' responses were in the question (8a) throughout the whole sample period. Figure 3.3.3 shows how firms on an aggregate level perceive or "nowcast" what has happened to the market in the previous quarter, the yellow line represents the official data corresponding to the firms' nowcasts.



As expected firms adapt to changes in the market which indicates that perceptions about the recent past are more accurate than views for the near future as an indicator of manufacturing output. As a result one would expect that nowcasts are more correlated with the outturn than forecasts are. The high percentage of firms reporting that the growth will remain the same makes sense because in the majority of the *in sample* period the growth is relative steady. In fact output growth time series during the *in sample* period turns out to be stationary after an augmented dickey fuller test for a unit root $p - value \ = \ 0.2081$ which indicates no evidence of a unit root. In Table B1 one can notice the magnitude of the financial crisis is captured by the ITS firms, this is result is evident when comparing the *in sample* and *out of sample* by looking at the *maximum* percentage of firms that reported "*Fall*" *in sample* 52% attributed to the negative peak of the early 1990 recession against that reported in the *out of sample* 59% attributed to the peak in the financial crisis 2009Q1. Similar results apply for the expectation series 43% against 50%.

This concludes the descriptive analysis. Next, is the *in sample* analysis where different quantification methods are applied on the expectation series (prospective data) in order to find the best model that explains the most variation between the alleged quantified expectation series and the official output growth. Afterwards the models will be assessed on their forecasting abilities during the *out of sample* crisis experiment of Section 4.

3.4 In sample analysis: Normal Times

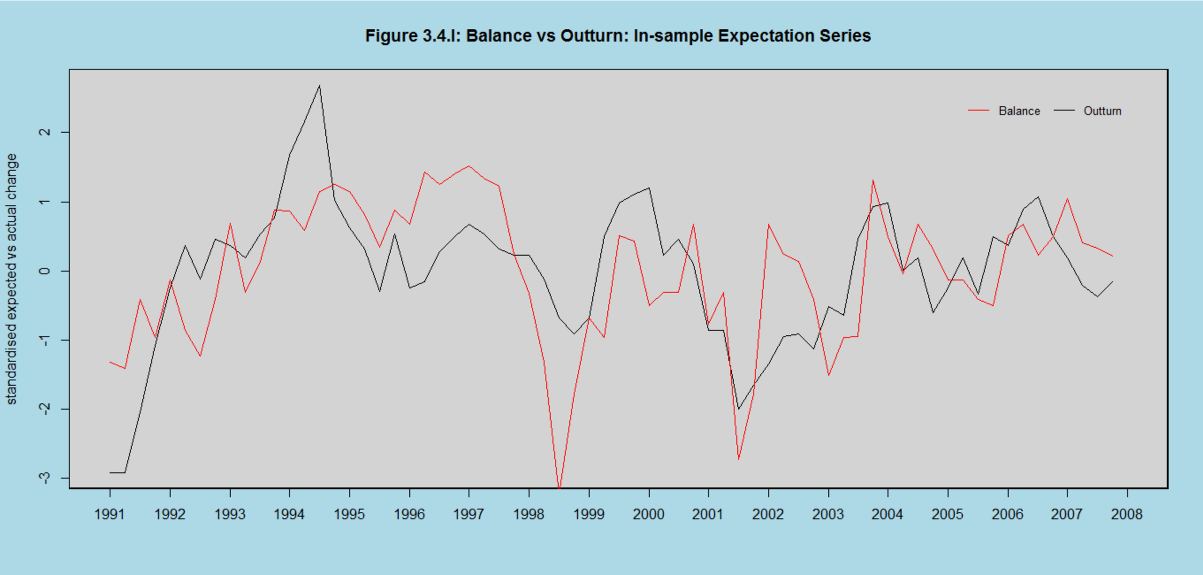
This part of the paper is the link between the quantification methods discussed in section 2 and the dataset as described in section 3 and Appendix Table B where a visual representation is outlined. The *in sample* analysis is based on the official data from 1991Q1 to 2008Q1. Now, because prospective data (expectation series) should give one quarter ahead *in sample* forecast have to be chosen from 1991Q1 to 2007Q4 and correspond to 1991Q2 to 2008Q1 official data. The 2008Q1 prospective views will not be used during the *in sample* analysis because they correspond to 2008Q2 (*out of sample*) official data which is not observed yet. The retrospective data give an early nowcast for the official data thus are chosen from 1991Q1 to 2008Q1. To provide one quarter ahead forecast for the *in sample* official data the quantification methods discussed in Section 2 were used and the results are summarised in Table B2.

Let's start by describing the procedure for each method starting with the Balance Statistic. The goal is to obtain a quantified proxy series x_t^e from the qualitative expectations and then examine the goodness of fit on the corresponding official data series (outturn).

I. Balance statistic

The indicator series from the Balance Statistic was called x_t^{BAL} and is computed as in (0) by connecting the expectations series (R^e, F^e) 1991Q1: 2007Q4 with the official data for output growth (x) 1991Q2:2008Q1 denoted as (y) in Appendix Table B. First $\hat{\theta} = 8.753$ is estimated and then $x_t^{BAL} = 8.753 \ (R_t^e - F_t^e)$ is calculated $\forall t \in [t_{1991\text{Q}1}, t_{2007\text{Q}4}]$. The result from the Balance quantification x_t^{BAL} is an one quarter ahead forecast that corresponds to x_t $\forall t \in$

$[t_{1991Q2}, t_{2008Q1}]$. The Figure 3.4.I shows how the expectation series quantified by the Balance statistic method capture the future movement of the output growth the correlation between x_t^{BAL} and x_t is 0.57 which is not that high. Notice that the plot uses the standardized version of the Balance statistic.



II. Probability method

The indicator series from the Carlson & Parkin (1975) approach was called x_t^{CP} and is computed as in (8) by connecting the expectations series (R^e, F^e) 1991Q1: 2007Q4 with the official data for output growth (y) 1991Q2:2008Q1. First the $\hat{\lambda} = 4.04$ is estimated and then x_t^{CP} is calculated. Because of evidence of non-normality form Table B1 and the lack of evidence to support the symmetry assumption of the indifference limen $[-4.04, 4.04]$, different extensions on (8) were implemented. First by changing the distribution function in to central $t(n = 6)$ with six degrees of freedom and $logistic(0,1)$ and secondly by allowing the indifference limen to be asymmetric $[-b, a\,]$. To estimate a and b a regression is used as $y_t \sim bX_t^b + aX_t^a + e$ where $X_t^b = \frac{f^e}{f^e - r^e}$ and $X_t^a = \frac{r^e}{f^e - r^e}$ and the parameters $\hat{a} = 5.38$,$\hat{b} =$ 5.05 are the OLS estimators. Then the new asymmetric indifference limen becomes $[-5.05, 5.38]$ which does not indicate significant asymmetry. Visually looking at the Figure 3.4.II and Table B2 it is evident enough to conclude that there is no substantial differences between the probability methods. Carlson and Parkin method seems robust enough even though normal and symmetric indifference interval assumptions are violated, x_t^{CP} does not seem to perform significantly worse from the other methods. In Table 3.4.II surprisingly enough the Balance Statistic which is the most restrictive quantification method seems to slightly outperform the less restrictive Probability methods in terms of correlation and RMSE. Seems by relaxing some of the assumptions of the Balance and CP method and allowing for asymmetry and heavier tails on the distribution does not give any advantage.

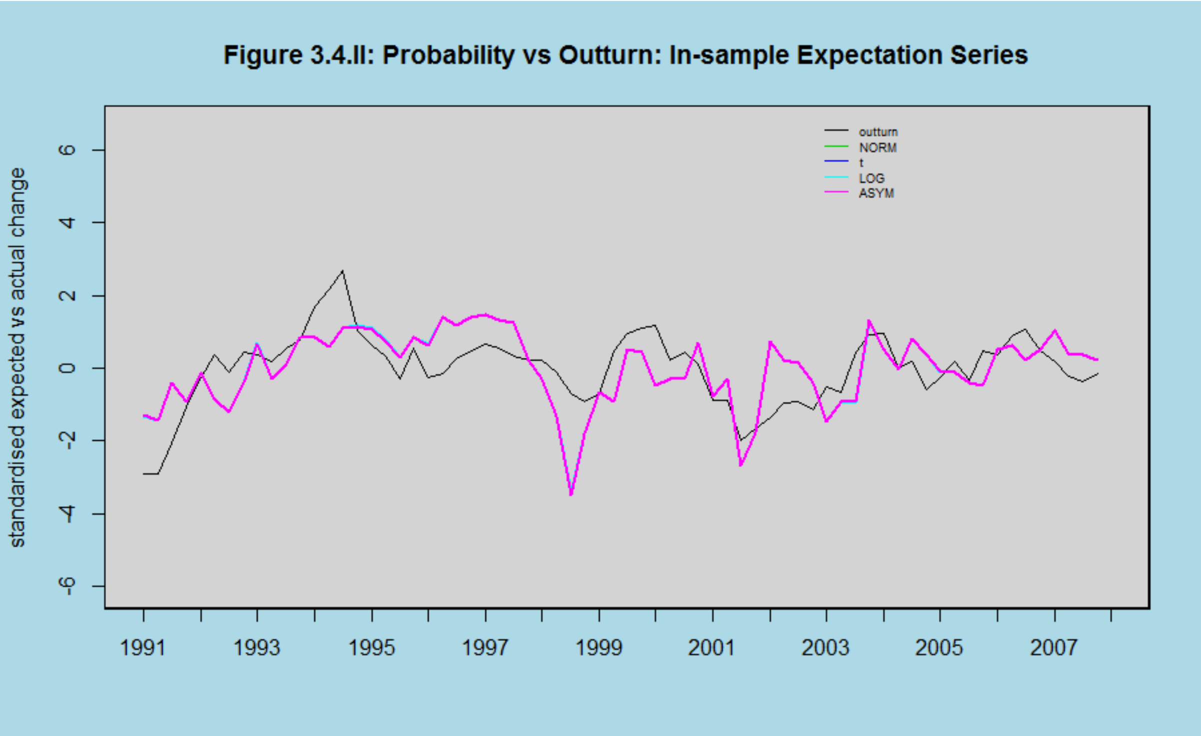


Table 3.4.II: *Probability and Balance method against the outturn* x_t		
	CORR	RMSE
x_t^{BAL}	0.570	1.858
x_t^{CP}	0.562	1.893
$x_t^{central-t}$	0.564	1.8924
$x_t^{logistic}$	0.5646	1.8923
x_t^{ASYM}	0.562	1.871

III. Regression Method

The regression method is quite challenging not only because, one has to effectively combine all the available data (retrospective, prospective and official) in order to derive an one quarter ahead forecast but also, one encounters many problems during the estimation procedure that could turn out difficult to be solved. The process to obtain the quantified expectation series involves two steps. To describe the steps Anderson's model (13) is used. In the first step (A) the regression model is estimated using the official data 1991Q1:2007Q4 as an independent variable and the retrospective data 1991Q1:2007Q4 as dependent variables. Then in (B) after the OLS coefficients $\hat{a}, \hat{b}$ are obtained from the regression (13) are plugged in (19) to obtain one quarter ahead forecast. The one period ahead forecast obtained from (19) s basically a proxy series for the official data 1991Q2:2008Q1. During the regression analysis of stages (A) and (B) of the model (13) many problems arise when testing the linear regression assumptions for the OLS estimators to hold. Some of the problems along with solutions are outlined below.

Step 1: Estimate the Anderson model (13) using the Retrospective data.

$$\textbf{A.}\ \ x_t = \underset{(1.236)}{13.9}R_t - \underset{(1.119)}{10.99}F_t + u_t$$

Step 2: Forecast by substituting $\hat{a} = 13.9$ and $-\hat{b} = 10.99$ in (19)

$$\textbf{B.}\ \ x_t^e = 13.9R_t^e - 10.99F_t^e$$

and obtain one step ahead forecast. One might be tempted to stop here. That would be fine if the linear regression assumptions would hold. If there is evidence of violation in the assumptions the OLS parameter estimations and their respective standard errors given by (A) do not hold anymore. Hence, the forecasts obtained from (B) will not be reliable anymore. The measurement error increases significantly. The linear hypothesis to be tested is for Normality, Autocorrelation, Heteroscedasticity and Multicollinearity. The regression model should at least show no evidence of serial autocorrelation in the residuals for someone to pass from (A) to (B) (see also Nardo 2003). If there is autocorrelation evident then $E[u]$ is not actual zero. To test for autocorrelation usually Durbin Watson (DW) statistic or a Box-Pierce test is used as well as plotting the autocorrelation function of residuals. For (A) the DW=0.97 and the Box-Pierce (p=0.0003) confirms the evidence on serial autocorrelation. Thus the model (A) has to be adjusted by allowing the error term to follow an autoregressive structure. GLS estimators are used as a fix for the correlation structure and the model (A) is re-estimated with an AR(1) structure on the error term.

$$\textbf{A2}\ x_t = \underset{(1.9)}{6.13}R_t - \underset{(1.8)}{5.37}F_t + u_t$$

$$u_t = 0.752\, u_{t-1} + w$$

Box-Pierce test: p-value = 0.7146, DW = 2.1, KS test: p-value = 0.4689

The model (A) was re-estimated in (A2) to adjust for the serial autocorrelations the parameters are considerably lower but not their respective standard errors. The new parameters $a^{GLS} = 6.13$ and $b^{GLS} = 5.37$ are the General Least Squares estimators. The Kolmogorov Smirnov (KS) is used to test the null hypothesis that w comes from a standard normal distribution since there is no evidence to reject that hypothesis w is treated as a normal. The Box-Pierce p.value and DW statistic indicate autocorrelation is fixed. The adjusted $\bar{R}^2$ is not reported because from

(A2) only the GLS parameters are needed to plug in (B2). The residual w is the independent identically distributed error term of (A2). The re-estimated model (A2) passed all the necessary diagnostics (see Appendix Table B2 and Figure B2) thus to obtain the adjusted for autocorrelations forecast (B) now becomes

$$\textbf{B2}\ y_t^e = 6.13R_t^e - 5.37F_t^e + 0.73\,u_{t-1}$$

During the analysis no regression models passed the diagnostic tests of stage A and B and had to be re-estimated as in (A2), to at least adjust for autocorrelation and then go to B2 to obtain the forecasts. Many regression models as extensions of (13) were estimated with different ARMA correlation structures on the residuals. The following five models [12]selected based on AIC and BIC criteria and the success on the diagnostic tests against their counterparts. For these five models one quarter ahead *in sample* forecasts are obtained using the prospective data. The best model *in sample*, will be the one that provides the best proxy for the official data. Keep in mind that model which performs best *in sample* does not mean it forecasts better *out of sample*.

The estimated regression models are summarised below[13] and their *in sample* evaluation in Table 3.4.III.

- Restricted Anderson Model with AR(1): (13) for $a = b = \lambda$

$$x_t = 6.2\,B_t + u_t$$

$$u_t = 0.759u_{t-1} + w$$

$$x_t^{RAM} = 6.2\,(R_t^e - F_t^e) + 0.759u_{t-1}$$

- Restricted Anderson Model version 2 (Thomas 1995)[14] with AR(1): (13) without R_t including an intercept c.

$$x_t = 3.74 - 12.8\,F_t + u_t$$

$$u_t = 0.722u_{t-1} + w$$

$$x_t^{RAM2} = 3.74 - 12.8F_t^e + 0.722u_{t-1}$$

- Unrestricted Anderson Model with AR(1): (13)

$$x_t = \underset{(1.9)}{6.13R_t} - \underset{(1.8)}{5.37F_t} + u_t$$

[12] The complete results from the in sample model selection and diagnostics was a result from a trial and error analysis. Models with including an intercept or excluding a variable and different ARMA structures for autocorrelation were considered and tested. The results are not presented here only a brief summary of the chosen five models. The reader can request to get the full results and code used for the analysis.

[13] The models presented here followed a similar two stage procedure as in A to A2 and then B2. All models were tested for diagnostics and found no evidence of serial autocorrelation after the adjustment. The parameters were found to be statistically significant for all cases except from the SMcA.

[14] The Restricted Anderson Model 2 $x_t = c - bF_t + u_t$ was used in order to avoid the multicollinearity problem that may occur because F_t and R_t in a sense give the same information. In Thomas (1995) the percentage of firms reporting a "*Fall*" found to be better predictor than using both.

$$u_t = 0.752\, u_{t-1} + w$$

$$x_t^{UAM} = 6.13R_t^e - 5.37F_t^e + 0.73\, u_{t-1}$$

- Pesaran Model with AR(1): (17)

$$x_t = \frac{11.06\, R_t - 9.07F_t + u_t}{1 - 0.71\, R_t}$$

$$u_t = 0.83\, u_{t-1} + w$$

$$x_t^{PES} = \frac{11.06\, R_t^e - 9.07F_t^e + 0.83\, u_{t-1}}{1 - 0.71R_t^e}$$

- Smith & McAleer Model with AR(1)[15]: (18)

$$x_t = \frac{8.25\, R_t - 6.65F_t + u_t}{1 - R_t - 0.57F_t}$$

$$u_t = 0.812\, u_{t-1} + w$$

$$x_t^{SMcA} = \frac{8.25\, R_t^e - 6.65F_t^e + 0.812\, u_{t-1}}{1 - R_t^e - 0.57F_t^e}$$

To assess the models on their *in sample* forecasting performance once again the correlation with the outturn will be used as well as the *in sample* RMSE[16]. Results are summarised in table

Table 3.4.III *Quantified expectations against the outturn* x_t

	CORR	RMSE
x_t^{RAM}	0.708	1.622
x_t^{RAM2}	0.706	1.775
x_t^{UAM}	0.711	1.621
x_t^{PES}	0.70	2.213
x_t^{SMcA}	0.64	3.724

From Table 3.5 the Unrestricted Anderson Model, (UAM), (13)(19) seems to have the most correlation with the outturn and the least measurement error. SMcA, (18)(21) seems to perform the worst *in sample*. Estimating the SMcA model before adjusting for autocorrelations no parameter is found to be significant, after adjusting for autocorrelations still the parameters are not found to be significant. This means that there is no need to extend the Pesaran's model and allow for an asymmetric relationship in both "*Rise*" and "*Fall*". Despite the *in sample* performance, SMcA is also used in the *out of sample* to investigate how it performs against the other quantification methods. Pesaran's model (17) is pretty close performance wise with the

15 For an even more extended version of SMcA model see Smith & McAleer (1995). In their paper they extend Pesaran's model (20) and propose (21). Then themselves estimate (21) as a time-varying parameters model which leads to the estimation of a nonlinear dynamic regression model.

16 RMSE is scale dependent thus to compare different aggregate indicators someone has to be sure they are in the same scale. Otherwise one can use NRMSE.

Anderson models and remains to be further tested in the *out of sample* forecast. The Restricted Anderson Model 2, (RAM2) utilizes only the percentage of firms reporting/expecting a "*Fall*" on the underlying output and seems to perform as well as the Balance (RAM) and also outperforms the Pesaran's model. This may indicate that firms' responses indicating a "*Fall*" contain more information than "*Rise*" about the underlying movement of output. Figure 3.4.IV is a visual representation of the indicator quantified using the Unrestricted Anderson's model. In Figure 3.4.V the *in sample* performance of all indicators obtained by the regression method is summarised.

Figure 3.4.IV: Unrestricted Anderson Model vs Outturn: In-sample Expectation Series

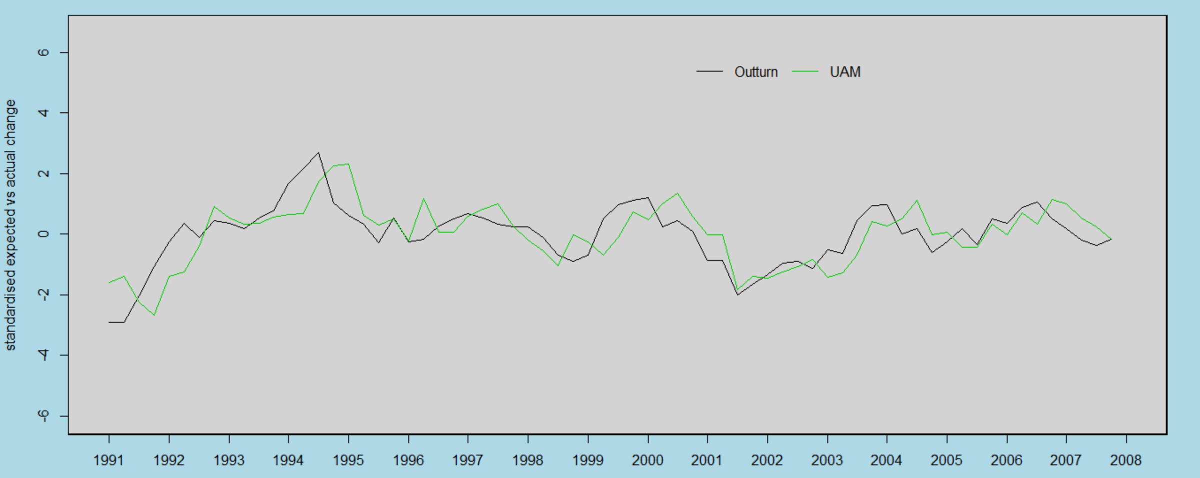
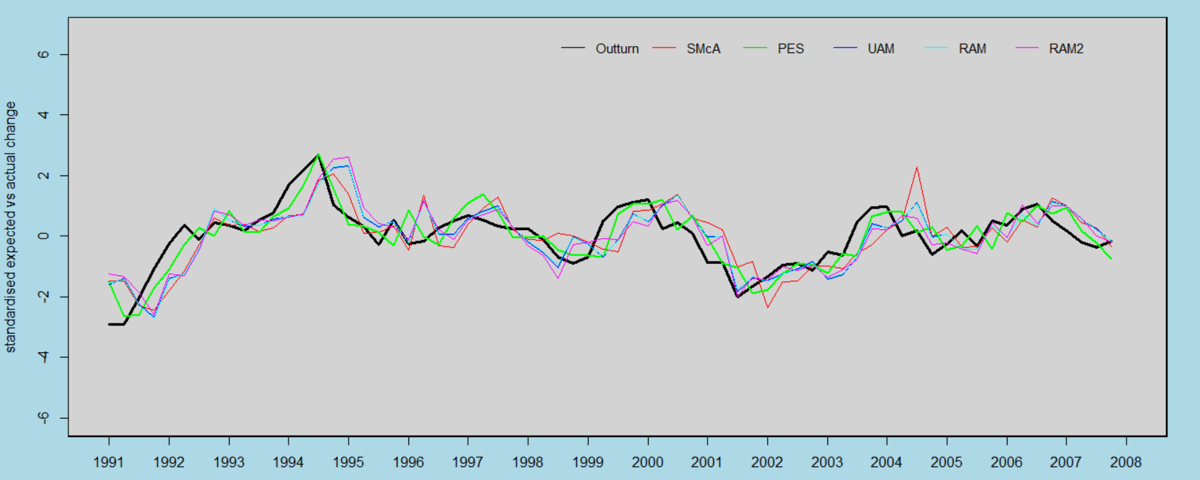


Figure 3.4.V: Regression Methods vs Outturn: In-sample Expectation Series



To conclude the *in sample* analysis from Tables 3.4II and 3.4III as it seems the majority of regression methods outperform the other methods. The best *in sample* fit is attributed to the Unrestricted Anderson Model (UAM) with AR(1) structure on the error term, for diagnostics and model summary go to Appendix Table B2 and Figure B2.

4. Out of sample analysis: The forecasting experiments

Following the *whole sample* and *in sample* analysis in section 3, it is time to test the indicators' performance on data that have not yet been observed. The *out of sample* analysis is conducted as a separate experiment from the *in sample* analysis of section 3. The *out of sample* period starts as the *in sample* period ends, from 2008Q2 and ends in 2009Q2, i.e. the height of the financial crisis and its aftermath. During the *in sample* analysis, the model parameters were estimated using all the historical information from 1991Q1 to 2007Q4 of both retrospective views and official data. By connecting the estimated parameters with the prospective data (1991Q1 to 2007Q4), a one quarter ahead forecast for 1991Q2 to 2008Q1 was obtained.

4.1 Real time forecasting from normal to crisis times

To begin the experiment, simply assume that an economist is standing on April 24 2008 and gets his hands on the just published Industrial Trends Survey. He wants to provide an indicator to track what is going to happen to manufacturing output from 2008Q1 to 2008Q2 using the quarterly ITS for April 2008 and the latest available data on manufacturing output. The official quarterly data for 2008Q1{Jan, Feb, Mar} are supposed to be published sometime in May 2008. As a result of that publication lag, the latest available quarterly data from ONS correspond to the previous quarter 2007 Q4 {Sep, Oct, Dec}[17]. In order to derive an indicator for 2008Q2 he will need the 2008Q1 official data (which are unobserved at the moment) and the ITS prospective views (responses to question **8b**) attributed to 2008Q2. To do that he will have to use the retrospective data to infer[18] 2008Q1. To quantify the 2008Q1 aggregate retrospective data, he will have to connect the (*in sample*) historical ITS retrospective data with the historical realisations of the official data, which means the period 1991Q1:2007Q4. So both retrospective and official from 1991Q1:2007Q4 will be used to estimate the models in a similar way to the *in sample* analysis. After the model estimation, the parameters from each quantification method are obtained. They are then related to the latest retrospective data (R_{08Q1}, F_{08Q1}) and a nowcast is obtained for the unobserved x_{08Q1} denoted as $\hat{x}_{08Q1}$. The models are then re-estimated for each method by connecting the retrospective data [1991Q1:2008Q1] with the official data [1991Q1:2007Q4, $\hat{x}_{08Q1}$]. After the models are estimated and the parameters are obtained for each, they are then applied to the latest ITS prospective data ($R^e_{08Q1}\ F^e_{08Q1}$) to get the one quarter ahead *out of sample* forecast for the output growth of 2008 Q2 denoted as $\hat{x}^e_{08Q2}$. [19] To obtain a forecast for the 2008Q3, the economist has to wait until the next quarterly ITS is

[17] In reality, ONS has published monthly quantitative official data for January and February of 2008. When the official data for March 2008 will be published by ONS in May 2008, it will be called 2008 Q1 official data for manufacturing output. Since the data available for the economist are assumed to be quarterly and in an aggregate form when published, then the only quarterly available data will consider to be those attributed to the previous quarter 2007 Q4{Sep, Oct, Dec} which ONS would have published in February 2008.

[18] The retrospective data attributed to 2008Q1 are the aggregate percentages of firms who answered question 8a: "*Up*", "*Down*" or "*Same*" in the ITS published on 24th April 2008.

[19] As an alternative, one can work straight on the expectations (prospective data) to get a proxy for 2008Q2. In order to infer 2008Q2 one has to estimate the models connecting the *in sample* prospective data $({}_{t-2}R^e_{t-1}, {}_{t-2}F^e_{t-1})$ from 1991Q2:2007Q3 and latest available official data (x_{t-1}) 1991Q1:2007Q4 obtain the parameters for each quantification method and the use them along with $({}_{t}R^e_{t+1}, {}_{t}F^e_{t+1}) = ({}_{08Q1}R^e_{08Q2}, {}_{08Q1}F^e_{08Q2})$ to obtain $\hat{x}^e_{08Q2}$ for 2008Q2 where the actual value of the outturn is x_{08Q2}. When dealing with prospective data one has to be very careful because it is necessary when agents form expectations at (t) for $(t+1)$ they all use the latest available official data x_{07Q4} that is the reason why ITS prospective data of 2007Q4 $({}_{t-1}R^e_{t}, {}_{t-2}F^e_{t}) = ({}_{07Q4}R^e_{08Q1}, {}_{07Q4}F^e_{08Q1})$ can not be used in the estimation process since in order to include them, x_{08Q1} should be included as well. The problem is that x_{08Q1} is assumed unobserved. But in the dataset as described in Table B during the calculations someone could include both by mistake this will definitely lead to overfitting .Because instead of estimating x_{08Q1} by nowcasting using the latest retrospective data $({}_{t-1}R_{t}\ , {}_{t-1}F_{t}\ )$ or surpassing x_{08Q1} (-1 obs) and go straight to forecast x_{08Q2}, the real value is used which means the overall measurement error will always be less or equal, intuitively $|x_{08Q1} - x_{08Q1}| = 0$ and $|x_{08Q1} - \hat{x}_{08Q1}| = u$.

available, which is July 2008. Moving from April to July, the official data for 2008Q1 would now be available, having been published in May 2008. As a result, to infer the output growth in 2008Q3, first the 2008Q2 must be estimated by connecting the historical retrospective data 1991Q1:2008Q1 with the historical official data 1991Q1:2008Q1 and the ITS July 2008 retrospective data in order to nowcast what has happened in 2008Q2 as $\hat{x}_{08Q2}$. The models are then re-estimated by connecting the retrospective data [1991Q1:2008Q2] and the official data [1991Q1:2008Q1,$\hat{x}_{08Q2}$] and obtaining the parameters from each quantification method, and, along with the ITS July 2008 prospective data, used to forecast output for 2008Q3. This recursive experiment is continued until the 2009Q1 prospective data are used to forecast the output of 2009Q2. The last forecast will involve estimating the models relating historical retrospective and official data from 1991Q1:2008Q4, with the latest retrospective data of 2009Q1 to nowcast 2009Q1 as $\hat{x}_{09Q1}$. The models are then re-estimated connecting retrospective data from [1991Q1:2008Q4] with official data from [1991Q1:2008Q4, $\hat{x}_{09Q1}$] to obtain the parameters for each quantification method. In the end, we use these parameters together with $({}_{09Q1}R^e_{09Q2}, {}_{09Q1}F^e_{09Q2})$ and obtain an one quarter ahead forecast denoted as $\hat{x}^e_{09Q2}$. After such a procedure, one should have 5 point estimations $\left[\hat{x}^e_{08Q2}, \hat{x}^e_{08Q3}, \hat{x}^e_{08Q4}, \hat{x}^e_{09Q1}, \hat{x}^e_{09Q2}\right]$ which are the quantified expectations corresponding to the *out of sample* crisis period. In order to assess their performance, one way is to measure the average error of prediction. The quantification method that produces on average the least amount of aggregate error is considered to perform the best. Thus the Mean Square Error (MSE) or the Root Mean Square Error (RMSE) is calculated for each method, just like in the *in sample* analysis.

(1) $MSE^k = \frac{1}{5}\sum_{i=1}^{5}\left({}_k\hat{x}^e_i - x_{t+i}\right)^2$ where $t = t_{2008\text{Q1}}$ and $k = 1,.N(m)$

(2) $RMSE^k = \sqrt{MSE^k}$

(3) $ME^k = \frac{1}{5}\sum_{i=1}^{5}\left({}_k\hat{x}^e_i - x_{t+i}\right)$

$N(m)$ is the number of quantification methods used in the experiment. The method which has min(MSE^k) or min($RMSE^k$) is considered to be the one that predicts output during the crisis the best. This type of assessment focuses on the magnitude of error not the direction (sign + $-$). Since the expectations data may contain all sorts of measurement errors already (such as sampling error, weights error, bias error, time publication error, lack of rationality etc.), it seems unfair to judge them by just their accuracy as seen in Table 4.1**.**

From Table 4.1.2, unexpectedly the Balance statistic shows the least measurement error. This result underlines the robustness of the Balance statistic and highlights why it is preferred by many institutions as an indicator. Although, all the methods perform to around the same standard on average, the Balance statistic is shown to steadily capture the magnitude of negative shocks during the crisis period. Pesaran's model seems to be the second choice as for 2009Q2 it manages to deliver the largest negative forecast -7.19% (on average) which was not that far behind what actually happened which -10.8%. To summarise, the quantified indicators from Table 4.1 correspond to the average value of forecast for each quarter. Table 4.1.2 outlines the RMSE (1) and ME (3), also one can notice the term $x_t^{ARMA(1,1)}$ which is an *out of sample* forecast using an ARMA(1,1) on past realisations of the output growth x_t . Encouragingly, the expectation measures obtained from all quantification methods outperform $x_t^{ARMA(1,1)}$. This result indicates once again the importance of the survey data as an early indicator. In other words, if one attempted to forecast the outturn of 2008Q2 using only quarterly data to do it, then standing in April 2008, the latest available data would be 2007Q4, which means by not having access to the ITS survey, one has to forecast 2 periods ahead. It turns out, despite the fact that the weighted aggregate prospective data contain a noticeable amount of error, they still manage to outperform statistical time series models that use only quantitative data on past values of output growth. This result suggests two points: first, a model which combines quantitative data and survey data such as retrospective and prospective views could produce more accurate forecasts and thus better indicators; second, that firms on an aggregate level seem to be able to recognise patterns in the economy and this result may indicate the rationality of firms when forming expectations.

Table 4.1: Average quantified measures for the financial crisis

	2008Q2	2008Q3	2008Q4	2009Q1	2009Q2
x_t^{BAL}	0.21	-1.88	-5.12	-5.57	-2.25
x_t^{CP}	0.06	-0.59	-3.77	-5.05	-3.44
x_t^{CP-t}	0.06	-0.59	-3.62	-4.74	-3.24
x_t^{ASYM}	-0.63	0.18	-2.78	-2.64	-4.45
x_t^{RAM}	-0.5	-0.08	-2.98	-3.34	-6.01
x_t^{RAM2}	-0.82	0.06	-3.1	-3.16	-5.19
x_t^{UAM}	-0.5	-0.08	-2.98	-3.34	-6.01
x_t^{PES}	-1.08	0.13	-2.67	-2.83	-7.19
x_t^{SMcA}	-1.26	0.46	-4.29	-2.94	-3.96
$x_t^{ARMA(1,1)}$	-0.14	0.28	-1.11	-1.9	-5.89
x_t	-1.5	-2.7	-7.2	-12.4	-10.8

Table 4.1.2: Measuring the accuracy of the predictions for the financial crisis period 2008Q1:2009Q2.

	ME	RMSE
x_t^{BAL}	-4	5.05
x_t^{CP}	-4.36	5.04
x_t^{CP-t}	-4.49	5.21
x_t^{ASYM}	-4.85	5.73
x_t^{RAM}	-4.34	5.11
x_t^{RAM2}	-4.48	5.32
x_t^{UAM}	-4.34	5.11
x_t^{PES}	-4.19	5.17
x_t^{SMcA}	-4.52	5.57
$x_t^{ARMA(1,1)}$	-5.17	6.04

The real-time forecasting experiment as described above does not give the full picture. One cannot still conclude that ARMA indicators are outperformed by Survey-based indicators both in normal times (here *in sample*) and crisis times (*out of sample*). To test that further, one has to test the robustness of survey data in both normal times and in times of crisis.

4.2 Evaluating indicators' performance in normal vs crisis periods

The evaluation of the predictive power of an indicator is a crucial step. Although, as seen in Table 4.1 and 4.2 the measurement error of prediction of Survey-based indicators is high, it is nonetheless significantly better than what one would get using ARMA models. Of course the measurement error is not the only important attribute: the correlation and the ability to move in the same direction as the official data is also of great interest when evaluating indicators. In the above forecasting experiment we saw that indicators calculated using historical data on normal periods outperform "naïve" forecasts only in the *out of sample* crisis period. The ideal scenario would be the one that in a parallel universe 2008Q2-2009Q2 would not be a crisis and instead occurred as normal times. Then if the survey data, trained, on the same *in sample* period 1992:Q2-2008Q1 outperformed ARMA forecasts in the parallel universe, then survey data should always forecast better. Thus, a better evaluation should be done in a simulation context where one can create that scenario and control comparing dissimilar real time scenarios, as in Claveria (2006) et al. Now in order to gather the necessary evidence, we have conducted three experiments to examine the performance both in normal and in crisis. To do that, it is necessary to compare how well survey data predict normal periods and crisis periods against ARMA models. For example, if the performance of survey data is better in both normal and crisis than ARMA models and survey data also perform similarly in normal times and crises, one can conclude that the crisis period does not have an effect on the performance of survey data.

The experiments are outlined below and the results from these experiments are in Appendix A Table 4A.

- **Experiment 1:** From Normal to Normal + Crisis
  - IN-SAMPLE : Normal Times[20] 1992Q1-2006Q4
  - OUT-OF-SAMPLE : Normal and Crisis[21] 2007Q1-2009Q2
- **Experiment 2:**
  a) From Normal to Normal
  - IN-SAMPLE : Normal Times 1992Q1-2006Q4
  - OUT-OF-SAMPLE : Normal Times 2007Q1-2008Q1

  b) From Normal to Crisis
  - IN-SAMPLE : Normal Times 1993Q2-2008Q1
  - OUT-OF-SAMPLE : Normal Times 2008Q1-2009Q2
- **Experiment 3:**
  - IN-SAMPLE : Normal Times 1992Q1-2003Q1
  - OUT-OF-SAMPLE : Normal and Crisis Times 2003Q2-2009Q2

[20] Normal period is considered a (weakly) stationary period (no unit root) or a period that is it not registered as a crisis. Crisis period is always considered the financial crisis (2008Q2-2009Q2) except if it is stated otherwise. The attempt to go back further and examine other crisis may or may not help to evaluate because many other factors (not taken into account) that have an effect on survey data change.

[21] The experiments are the same for the arma models and survey-based models. Survey data with focus the on the expectations are evaluated both in comparison with arma forecasts but also with themselves between normal and crisis. Because the data are limited, only 5 quarterly crisis observations, the crisis period cannot be used for model estimation (*in sample*). Thus the out of sample will always contain the crisis period. The difference will be on whether one attempts to predict (out of sample) both normal and crisis or predicts normal and then

After these three experiments, we conclude that survey data perform differently in normal times and crisis times. This is evident from Table4A just by looking at the higher *out of sample* RMSE between normal and crisis as well as the significant difference between the *in sample* and *out of sample* RMSE in all experiments. Furthermore, all the quantified survey-based indicators outperform the ARMA models when in crisis, but the same is not true in normal times. First of all, the quantification method has a significant impact on the performance of survey-based indicators. Each method does not perform the same, however, as, for example, even if the survey views have a high predictive performance regarding the movement of the output, it does not mean the quantification method will be able to project it. Always at least one of the methods used outperforms the ARMA models. The best indicators are from the UAM method.

The results also suggest firms tend to respond and adjust their expectations when it is more vital to do so. When the period is smooth and the economy stable, the firms seem to rely more on past values of output, but when the shocks are high enough (crisis), they re-adjust immediately. Looking again at Figures 3.4II and 3.4III, the high percentage of firms reporting no growth for relatively small changes in output can be explained from the above result and the fact that in normal times, business is running smoothly and firms do not seem eager to predict a small downward or upward movement because they regard it as being "normal". During and after the financial crisis, what is considered "normal" is shifted more towards 1-2% change of movements in output which can be confirmed by the ITS APS (1998, 2008, 2013) see Appendix A Figure 4A.

4.3 Effects of sample size on survey data in normal vs crisis

Usually, in a severe crisis, the quality of survey data is threatened because many firms may withhold information or are too busy running their company and do not respond or even go out-of-business. Thus, it is important to examine whether changes in the performance of indicators are affected by changes in sample size during normal and crisis periods. To do that, experiments were conducted, specifically focused on what happened before and during the financial crisis period, with the goal being to determine how changes in sample size might affect performance. Because of the lack of observations, to do the experiments the dataset is changed to monthly data (3month on 3month growth). The new dataset is extended on the crisis period and contains 18 observations of what is considered normal times from Oct-2006 to Mar-2008 and another 18 observations of crisis times from Apr-2008 to Sep-2009. We examined whether the change in sample size is significant between two groups denoted "normal" and "crisis". This was done by creating a dummy variable called "period" and implementing a paired t-test [22]between the 2 sub periods (normal and crisis).

Thus we examined the hypothesis $H_0\colon \mu_{normal} - \mu_{crisis} = 20\ vs\ H_1\colon greater\ than\ 20$ the mean change in sample size from normal to crisis period is less than 5%. The results show no evidence to reject that hypothesis as the $p.value = 0.99$ which means there is no significant change in sample size before and during the crisis. Now we proceed one step further and examine what happened after the crisis (ex-post) by comparing the crisis period Apr-2008 to

[22] The paired t-test assumes that "between" samples normal vs crisis are dependent but "within" it assumes that the observations are independent with each other. Usually this is not a good test for time series observations. To examine the "within" autocorrelation we run two AR(1), one in normal period and one in crisis to see if there is autocorrelation within them. In normal period the AR(1) coefficient is 0.06% and in crisis is -0.24% which indicates a small negative correlation (not of much significance). Thus this results should be treated carefully, we need further evidence.

Sep-2009 against the ex-post period from Oct-2009 to Mar-2011. Following the same procedure, we test the hypothesis that the average change in sample between these two periods was more than 5%. Indeed, the results show with 95% confidence level the mean change was more than 5% $p.value = 4.632e-05$. Also the 10% level of change is also tested with a $p.value = 0.029$ and this means with 95% confidence that the effective sample size changed more than 10% before and after the crisis (for more detailed results see Appendix A Table 4A2). We found that the changes in sample size were not significant between ex-ante and during the crisis, but were significant ex-post. This result indicates the robustness of ITS survey during the crisis, but also suggests that the performance of the indicators from normal times to crisis is not affected by the small changes in sample size (ex-ante and during crisis).

To gather more evidence to support the claim that the effect of change in sample size from normal times to crisis times did not affect the performance of survey data, regression and anova experiments were implemented. The idea is to examine if the change in sample size between normal and crisis has any effect on the ability of the balance statistic[23] to predict the official data. After many trial and error experiments we found no evidence that the sample size has any effect on the regression. Actually, not only is the effect of change in sample size not significant, but also its interactions with the balance statistic and the period (binary: normal or crisis). The regression results are outlined in Appendix A Table 4A3.

The presence of a significant interaction indicates that the effect of one predictor variable (Balance Statistic) on the response variable (output growth) is different at different values of the other predictor variable(s) (sample size). It is tested by adding a term to the model in which the two predictor variables are multiplied. In fact, we tested two effects of sample size one with the balance and one with the period. The regression equation with all possible interactions will simply looks like this:

Model 1 $output \sim balance + balance * sample.size + period * sample.size$

The first part of the Model 1 becomes the basis of comparison between models and is very similar to the Restricted Anderson Model.

Model 2 $output \sim balance$

In order to test whether sample size has any significant effect we compared different model combinations while adding some or all of the sample size effects against the model with only the Balance Statistic **Model 2**. The testing was done via anova F-tests and AIC criteria. Also a backward elimination procedure is implemented on the unrestricted model with all possible interactions namely:

Model 0 $output \sim balance * sample * period$

This model has 7 parameters estimated. The backward elimination resulted to

$$output \sim balance + period + balance * period$$

which somewhat confirms previous results of statistically significant difference in performance of survey data between normal times and crisis, although the driver behind this difference in

[23] One can use any quantification method instead of the balance statistic. In this study the performance of the balance statistic is shown to be good enough and because its' use is very common amongst institutions it is the only one presented here. CP method gives similar results.

performance on normal vs crisis does not seem to be caused by the change in sample size of the ITS. The extra evidence we got from the regression experiments along with the t.tests indicate that the sample size and its interactions were never significant. One can also conclude that crisis is not the reason that firms do not respond in the surveys.

On the above regression models, we basically examined the effects that sample size has on output and not on the balance statistic. Our thinking was that if changes in sample size in normal times versus crisis are not significant, it follows that taking into account that effect does not improve the forecast performance of the model that has only the balance statistic. In other words, if **Model 2** prevails against models that have the change in sample size and its interactions with period, it means that knowing how the sample size changes through normal times and crisis does not improve the forecasting performance of the model depending only on the balance statistic.

We are more compelled to know whether sample size has any effect on the quality of the balance statistic. To do that, we need to find a measure to evaluate the quality of survey data. That measure is the ability to forecast manufacturing output. Thus, we will have to investigate the forecasting error produced from balance vs output and also take into account the different time periods normal vs crisis. Hence we examined the behaviour of the forecast residual term between survey expectations and manufacturing output, namely $e_{t+1} = x_{t+1} - {}_{t}B_{t+1}$. The idea is to investigate if by taking into account the changes in sample size between normal times and crisis times it can help predict any systematic pattern in the forecast error. In other words, if the effect of the sample size is different in normal versus crisis, then we should uncover a systematic effect in the error term. Also, if the change in sample size is significant during the whole period normal and crisis the effect of sample size alone should be significant.

Thus we start with the following model:

M: $forecast.error \sim intercept + sample.size + period + sample.size.period$

and we run a backward elimination process and we finally end up with the

M2:$forecast.error \sim intercept + \ period$

This supports what we already observed previously in the forecasting experiments, the $period$ plays a significant role in the nature of forecast errors and that is why we found that survey-based aggregate indicators perform differently in normal times compared to crisis. We also tried the same process without the intercept and we ended up with a model without coefficients which again does not include any effect of the $sample.size$.

Conclusion

The task of the paper was to provide the reader with tools to be able to connect and understand the theoretical procedures with a practical application and assess the performance of survey data in times of crisis. First, by identifying any significant differences regarding the performance of data between in normal and crisis times. Secondly, trying to identify the drivers of differences in performance between these two periods. Possible drivers that were examined were the sensitivity of quantification methods to capture temporary and permanent shocks, changes in sample size and changes in answering practices. We continue with a brief discussion on many controversial measurement errors that are entailed in the survey data that in certain time periods could give false positive results.

As one reads through the literature on quantification of survey expectations, one will encounter many contradictory results, mainly because of the fact that the quantified expectations from the methods discussed in section 2 and implemented in section 3 and 4 are more of an approximation of firms' expectations x_{it}^{e} rather the underlying economic variable x_t . Hence, one measurement error could come from the inability of the firm to correctly forecast. Usually, weights are used to compensate for the "trust" in predictions: a larger firm usually has a larger weight in the sample, which means if these firms poorly forecast or do not give an honest opinion in certain periods, it affects the quality of survey data significantly, especially on the aggregate form even if some of the (small size) firms forecast perfectly. Some authors like Mitchell (2002) et al. find that the unweighted data outperform the weighted data and suggest a panel data analysis to model each firm separately through time.

Another problem discussed briefly in section 3.2 is the matching process in order to ensure that survey data match the official data. It is impossible to find a perfect match between the survey and the official data, but the Industrial Trends Survey gives a high quality match when the manufacturing sector is concerned (see also Thomas (1995) and Mitchell (2010) et al.). The ITS is commonly used by many authors in the literature which means they appreciate the quality of the survey. Anyhow, the difficulty of finding a perfect match is evident and the error of matching aggregate survey data series with official figures going back 20-30 years is inevitable, although our answering practices through the years (1998, 2008, 2013) indicate an improving relationship between ITS and MPI.

There is also an error coming from the lack of rationality of firms when forming expectations. In our case many tests for unbiasedness, serial correlation and weak efficiency were implemented, but the results were contradictory between methods and periods and therefore were not presented. In a statistical sense, we have no evidence to conclude whether firms are rational on an aggregate level, but we have evidence that they can significantly and persistently beat forecasts that only depend on information from past values. That alone is an indicator that agents might be forming expectations that are rational, but overall it is not sufficient to conclude. As far as the sample bias error is concerned, it was tested, but only in the sense of measuring the effect of changes in the sample size from normal to crisis periods and the evidence showed that there is no significant effect of the change in sample that relates to the performance of survey data.

All these different measurement errors that are incorporated in the expectational data and the fact that survey expectational data basically are not a proxy for the underlying variable (output) but more of the true expectation values of firms, suggests that one should first test the models' ability to fit the true values of firms expectations and then attempt to evaluate the forecasting performance of survey on the aggregate macroeconomic variable and eventually test the difference in performance between normal and crisis times.

Of course there are no firm-level expectational quantitative data available, but the mechanism to generate them, exists, as outlined in the probability method section 2.2. For example, if someone were to evaluate between quantification methods, they could try a simulation experiment such as that described in Claveria (2006) et al. in order to better identify the size and the systematic nature of the error coming from the quantification methods and after defining that error, they should go on to assess how the data perform in normal times versus crisis times.

Returning to the analysis of sections 3 and 4 from the paper, the results confirm the utility and robustness of the use of Balance (BAL) (0) statistics by many institutions, as a figure to provide an early indicator and identify the upcoming shocks in the market. The Carlson & Parkin (CP) method (8) posed robustness even though none of the initial assumptions made to construct the model were evident. CP still performed relatively close to its counterpart extensions (t, asymmetric limen, logistic) as well as against the regression methods. As far as the regression methods are concerned, the Unrestricted Anderson Model (13) was found to fit the official data best when *in-sample* and *out of sample*. Pesaran's model (17) performed very close to the BAL and UAM, but it did not show enough evidence to be a clear choice of quantification for expectations thus in later experiments was omitted.

The regression method after all the experiments seems to outperform the other methods and the arma models overall. A noticeable attribute of regression as a conversion method is that one can combine many different sources of data (quantitative and qualitative) and try to find the best *in sample* fit and then attempt to forecast the future. This means that the survey data are being assessed as whole (both retrospective and prospective) and the fact that a combination of those two alone give the best results in the experiments adds value to the survey data themselves.

To conclude the discussion, the *out of sample* analysis was conducted as a progressive forecasting experiment on the financial crisis period of 2008Q2:2009Q2. Even though the magnitude of the crisis, as expected, could not be captured the results suggest, the quantified average measures from all quantification methods favorably outperform an ARMA(1,1) and AR(1), AR(2) in the latest experiments. This finding suggests that the utility of qualitative survey data is very high even in a severe crisis, not because of their forecasting abilities alone, but also because they can increase the forecasting abilities of other models when they are included in the model compared with when they are excluded (see section 4). The issue of quantifying survey-based expectations is very important, not only because it is necessary to find the best quantification method that provides the best forecast in terms of Table 4.1.2 and Appendix A Table 4A, but also the method that transforms more accurately the agents' true expectations into quantified expectation series. This is because one can add the quantified expectation series as predictors in other known macroeconomic models to increase their forecasting abilities. Also it is necessary to identify the size and systematic nature of the error coming from the conversion procedure in order to assess effects of other possible driving factors on the performance of survey data between normal and crisis periods.

To figure out if there is any impact from a crisis on survey data, we added a series of experiments and the results outlined in Appendix A Figure 4A4 suggest that there is a change in performance before and during the crisis. Of course, there could be a number of possible factors driving that change, including the quantification method. Thus, we examined many different conversion methods for qualitative expectations in order to find any significant differences between them in normal versus crisis times. Results from Tables 4.1, 4.2 and 4.A show that all quantification methods perform similarly but outperform "naive" forecasts. This basically means the driving factor behind the change in performance between normal and crisis is not the conversion methods, because if it was, we should have different results between them.

Having put aside the quantification methods, we then focused on the impact of change in sample size on survey data in normal versus crisis times. The results from Figures 4A2, 4A3, 4A4 indicate that the sample size does not play any significant role whatsoever in the difference in performance of survey data between normal and crisis periods. The robustness of the ITS during the crisis period is evident.

The ITS APS (1998, 2008, 2013) showed that during the crisis, firms were more conservative, fearing for a continuous downward trend. Replying "same" for output movements between 4-8% has significantly decreased compared to before (1998) and after (2013) the crisis period. The rise in % of firms classifying 0-2% as "the same" is smaller than the fall in those deeming 2-4% "the same" which could indicate one an improved relationship between the ONS and ITS data over time.

Although we did not identify exactly the drivers behind the difference in performance between normal and crisis period, we did not find evidence that the changes in sample size or the quantification methods or any change in answering practices are the real drivers behind it. Hence we are assessing the utility of survey data in their overall performance.

Finally, another compelling feature of the survey data comes from the fact that when the Industrial Trends Survey is published, an economist can use the "early views" of the firms and attempt to provide two indicators: one as a nowcast for the current period and one as a forecast for the next period. Combine this with the fact that the survey data may raise the forecasting ability of other macroeconomic models makes the utility of the survey data undeniable both in normal times and in times of crisis. Regardless of how well they perform, it is always useful to have them at your disposal before the official figures, rather than to not have them at all. For the future, it would be interesting to look at disaggregated analysis and weighted versus unweighted data, plus a more rigorous assessment of the rationale of firms using a wider range of quantitative variables as information for agents when forming expectations.

References

Anderson, O. (1952), 'The Business Test of the IFO-Institute for Economic Research, Munich, and its Theoretical Model', *Review of the International Statistical Institute* **20**, 1–17.

Batchelor, R. (1981), 'Aggregate expectations under the stable laws**',** *Journal of Econometrics*, **16**, p.199-210.

Batchelor, R. (1982), 'Expectations, Output and Inflation', *European Economic Review* **17**, 1–25.

Batchelor, R.A. (1984), 'Quantitative vs. qualitative measures of inflation expectations', *Oxford Bulletin of Economics and Statistics* **48** (2), 99–120.

Batchelor, R. & Orr, A. (1988), 'Inflation Expectations Revisited', *Economica* **55**, 317–331

Batchelor R. (2010), 'How robust are quantified survey data? Evidence from the United States**'** in Peter Sinclair (ed.), *Inflation Expectations* (by invitation), Oxford and New York: Routledge, ISBN 0-415-56174-4.

Berk, J.M. (1999) 'Measuring inflation expectations: a survey data approach', *Applied Economics,* 31:11, 1467-1480, DOI: 10.1080/00368499323337

Carlson, J. & Parkin, M. (1975), 'Inflation Expectations', *Economica* **42**, 123–138.

Claveria, O., Pons, E., Surinach, J. (2006) 'Quantification of Expectations. Are They Useful for Forecasting Inflation?' *Economic Issues*, Vol. **11**, Part 2, 2006

Driver, C. & Urga, G. (2004), 'Transforming Qualitative Survey Data: Performance Comparisons for the UK', *Oxford Bulletin of Economics and Statistics* **66**, 71–90.

Enrico D'Elia, (2005). 'Using the results of qualitative surveys in quantitative analysis', ISAE Working Papers **56**, ISTAT - Italian National Institute of Statistics - (Rome, ITALY).

Ece, O. (2013), 'Consumer Tendency Survey Based Inflation Expectations', Central Bank of the Republic of Turkey, Istiklal Cad. No:10, Ankara, Turkey, 06100, Working Paper NO:13/08

Lahiri, K. & Zhao, Y. (2013). 'Quantifying Heterogeneous Survey Expectations: The Carlson-Parkin Method Revisited'. *Discussion Papers* 13-08, University at Albany, SUNY, Department of Economics.

Mitchell, J., Smith, R. & Weale, M. (2002), 'Quantification of Qualitative Firm-level Survey Data', *Economic Journal* **112**, C117–C135.

Mitchell, J., Mouratidis, K., Weale, M., (2004), 'The impact of survey aggregation methods on the quality of business survey indicators', ECFIN/2003/A3-04.

Mitchell, J., Smith, R. & Weale, M. (2005), 'Forecasting Manufacturing Output Growth using Firm-level Survey Data', Manchester School **73**, 479–499.

Mitchell, J., Mouratidis, K., Weale, M., (2007), ‘Uncertainty in manufacturing: evidence from qualitative survey data’, Economics Letters, Vol. **94**, pp. 245-252

Lui, S., Mitchell, J. and Weale, M. (2010), ‘Qualitative business surveys: signal or noise?’ *Journal of the Royal Statistical Society : Series A (Statistics in Society),* Vol.**174**, No.2, 327-348.

Muth, J. (1961), ‘Rational Expectations and the Theory of Price Movements’, *Econometrica **29**, 315–335.*

Nardo, M. (2003), ‘The Quantification of Qualitative Survey Data: a Critical Assessment’, *Journal of Economic Surveys* **17**, 645–668.

Pesaran, M.H. (1984), Expectations formation and macroeconomic modelling, *in* P. Ma-grange and P. Muet, ed., ‘Contemporary Macroeconomic Modelling’, Blackwell, Oxford, pp. 27–53

Pesaran, M.H. (1985), ‘Formation of Inflation Expectations in British Manufacturing Industries’, *Economic Journal* **95**, 948–975

Pesaran, M.H. (1987), The Limits to Rational Expectations, Basil Blackwell., Oxford

Pesaran, M.H. & Weale, M. (2006), ‘Survey Expectations’. In G. Elliot, C. Granger, & A. Timmerman, Handbook of Economic Forecasting, Vol 1 (pp. 715-776). Elsevier.

Smith, J. & McAleer, M. (1995), ‘Alternative procedures for converting qualitative response data to quantitative expectations: an application to Australian manufacturing’, *Journal of Applied Econometrics* **10**, 165–185.

Steffen Henzel & Timo Wollmershäuser, (2005). ‘Quantifying Inflation Expectations with the Carlson-Parking Method: A Survey-based Determination of the Just Noticeable Difference’, *Journal of Business Cycle Measurement and Analysis* **2**, OECD Publishing, Centre for International Research on Economic Tendency Surveys, vol. 2005(3), pages 321-352.

Theil, H. (1952), ‘On the Time Shape of Economic Microvariables and the Munich Business Test’, *Revue de l’Institute International de Statistique* **20**.

Thomas, D.G. (1995), ‘Output expectations within manufacturing industry’, *Applied Economics*, **27:5**, 403-408. DOI: 10.1080/00036849500000145

Wren-Lewis, S. (1985), ‘The Quantification of Survey Data on Expectations’, *National Institute Economic Review* **113**, 39–49.

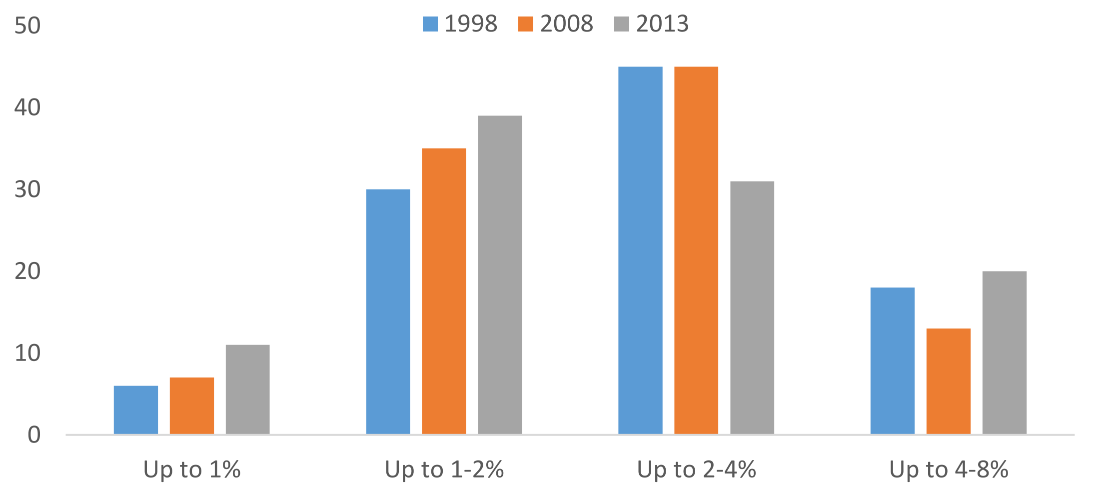
Appendix A

TABLE 4A	EXPERIMENT NO1					
NORMAL Forecast Both NORMAL + CRISIS Times	CORRELATION			RMSE		
	IN SAMPLE (59 obs)	OUT OF SAMPLE (10 obs)		IN SAMPLE (59 obs)	OUT OF SAMPLE (10 obs)	
DIFFERENT AGGREGATE SURVEY BASED INDCIATORS	Normal Times 1992Q1-2006Q4 (59 obs)	Normal Times 2007Q1-2008Q1 (5 obs)	Crisis Times 2008Q1-2009Q2 (5 obs)	Normal Times 1992Q1-2006Q4	Normal Times 2007Q1-2008Q (5 obs)	Crisis Times 2008Q1-2009Q2 (5 obs)
BAL	0.549	0.514	0.970	1.678	0.817	7.86
CP	0.541	0.486	0.961	1.708	1.16	4.47
RAM	0.799	0.791	0.845	1.194	1.34	5.24
UAM	0.802	0.792	0.842	1.228	1.26	5.18
ARMA(1,1)	0.779	0.754	0.702	1.216	0.889	6.39
AR(1)	0.777	0.727	0.708	1.221	0.936	6.36
AR(2)	0.783	0.775	0.699	1.214	0.847	6.40

EXPERIMENT NO2				
NORMAL Forecast NORMAL Times	CORRELATION		RMSE	
	IN SAMPLE	OUT OF SAMPLE	IN SAMPLE	OUT OF SAMPLE
INDICATORS	Normal Times 1992Q1-2006Q4	Normal Times 2007Q1-2008Q1	Normal Times 1992Q1-2006Q4	Normal Times 2007Q1-2008Q1
BAL	0.549	0.578	1.678	1.447
CP	0.541	0.550	1.708	1.031
RAM	0.798	0.588	1.205	0.994
UAM	0.791	0.600	1.271	0.977
ARMA(1,1)	0.779	0.311	1.216	1.296
AR(1)	0.777	0.282	1.221	1.337
AR(2)	0.783	0.342	1.214	1.261
NORMAL Forecast CRISIS Times	CORRELATION		RMSE	
	IN SAMPLE	OUT OF SAMPLE	IN SAMPLE	OUT OF SAMPLE
INDICATORS	Normal Times 1993Q2-2008Q1	Crisis Times 2008Q1-2009Q2	Normal Times 1993Q2-2008Q1	Crisis Times 2008Q1-2009Q2

BAL	0.54	0.966			1.64	7.870		
CP	0.53	0.957			1.65	4.669		
RAM	0.81	0.857			1.22	5.215		
UAM	0.81	0.852			1.28	5.129		
ARMA(1,1)	0.79	0.705			1.19	6.363		
AR(1)	0.79	0.711			1.19	6.338		
AR(2)	0.78	0.702			1.16	6.374		
EXPERIMENT NO3								
NORMAL Forecast Both NORMAL + CRISIS Times	CORRELATION				RMSE			
	IN SAMPLE	OUT OF SAMPLE			IN SAMPLE	OUT OF SAMPLE		
INDICATORS	Normal Times 1992Q1-2003Q1	Normal Times 2003Q2-2008Q1	Crisis Times 2008Q2-2009Q2	Both 2003Q2-2009Q2	Normal Times 1992Q1-2003Q1	Normal Times 2003Q2-2008Q1	Crisis Times 2008Q2-2009Q2	Both 2003Q2-2009Q2
BAL	0.547	0.583	0.267	0.807	1.797	1.447	7.912	3.768
CP	0.540	0.554	0.191	0.783	1.827	1.025	5.644	2.685
RAM	0.838	0.585	0.847	0.923	1.198	1.001	4.083	2.034
UAM	0.829	0.598	0.898	0.932	1.276	0.978	4.359	2.136
ARMA(1,1)	0.815	0.311	0.703	0.722	1.222	1.296	6.393	3.085
AR(1)	0.811	0.282	0.709	0.716	1.235	1.337	6.363	3.087
AR(2)	0.822	0.342	0.699	0.727	1.214	1.261	6.402	3.078

Figure 4A:
ITS APS: what range of movement would you regard as falling within the reply "the same"?

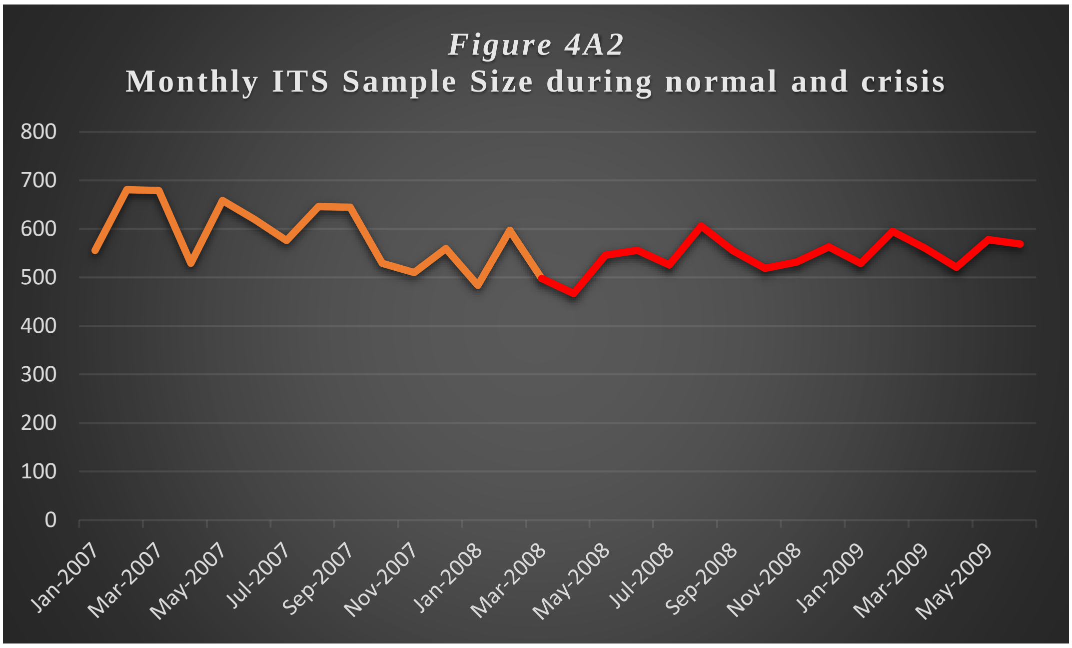


Q9b on the ITS ASP asks what range of movement would the respondent classify as “the same”, when responding on output. The distribution of these movements seems to have flattened over time – with more respondents classifying a wider range of movements as “the same” relative to 1998.

During the early stages of financial crisis in April 2008 firms responded to the ITS Answering Practice Survey. The % of firms classifying a change of 4-8% in output “the same” has decreased also the %0-2 has increased compared to 1998 this might be an indicator that more firms recognise the start of the crisis -2% in 2008Q2. The APS of 2013 show an uptick in % of firms classifying movements of 0-2% as “the same”, and a reduction in those classifying movements of up to 2-4% in this category. The latter is encouraging, as it suggests that larger changes in output are being captured in the ITS data.

The increase in smaller movements being classified as “the same” may mean that more incremental changes in output aren’t be captured by the output balance.

But the *rise* in those classifying 0-2% as “the same” is smaller than the *fall* in those deeming 2-4% “the same” – so on net, one might expect this to have improved the relationship between the ONS and ITS data over time. Although given the persistent flattening, this might now have anything to do with the financial crisis.

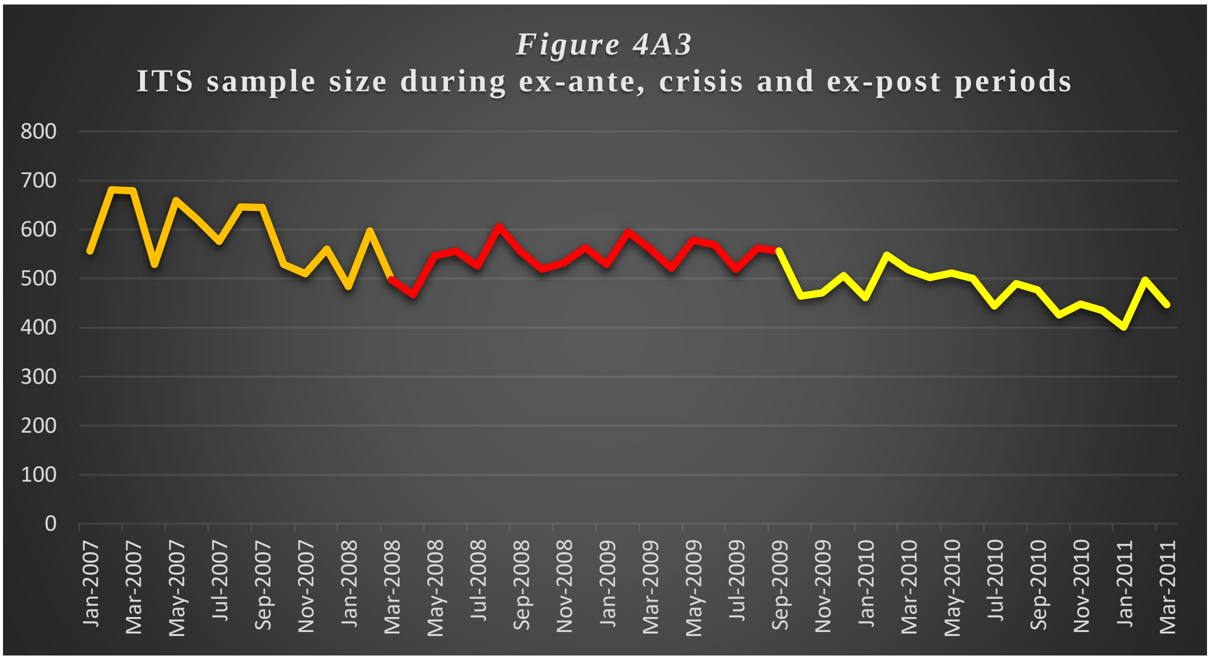


Followed by the procedure of testing if changes in sample size between normal and crisis times are of any significance.

```
> # TEST FOR EQUAL VARIANCES
> # Ho: Variance ratio between samples is 1
> # Performs an F test to compare the variances of two samples from normal populations.
>
> a = sample.size[period=="normal"]
> b = sample.size[period=="crisis"]
> var.test(a,b)

        F test to compare two variances

data:  a and b
F = 3.7726, num df = 17, denom df = 17, p-value = 0.009098
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  1.41121 10.08526
sample estimates:
ratio of variances
          3.772589
```

```
> t.test(sample.size~period,mu = 20 ,alternative ="greater",var.equal = FALSE,paired=TRUE)

        Paired t-test

data:  sample.size by period
t = -4.324, df = 17, p-value = 0.9998
alternative hypothesis: true difference in means is greater than 20
95 percent confidence interval:
 -59.6984      Inf
sample estimates:
mean of the differences
              -36.83333
```

Now we extend the period to examine if ex-post there is any significant change in the ITS sample size.



Followed by the procedure of testing if changes in sample size between crisis and ex-post times are of any significance.

```
> # TEST FOR NORMAL DISTRIBUTIONS
> # Ho: Normal
> shapiro.test(sample.size)

        Shapiro-Wilk normality test

data:  sample.size
W = 0.97906, p-value = 0.7136

> shapiro.test(sample.size[period==0])

        Shapiro-Wilk normality test

data:  sample.size[period == 0]
W = 0.95445, p-value = 0.4989

> shapiro.test(sample.size[period==1])

        Shapiro-Wilk normality test

data:  sample.size[period == 1]
W = 0.98509, p-value = 0.9872

> # TEST FOR EQUAL VARIANCES
> # Ho: Variance ratio between samples is 1
> # Performs an F test to compare the variances of two samples from normal populations.
> var.test(a,b)

        F test to compare two variances

data:  a and b
F = 0.74656, num df = 17, denom df = 17, p-value = 0.5534
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2792668 1.9957902
sample estimates:
ratio of variances
         0.7465641

> # Dependent or Independent samples? => Dependent
>
> mean(a)*0.05
[1] 27.38889
>
> t.test( a,b,m=27,alternative ="greater",paired=TRUE,var.equal=TRUE)

        Paired t-test

data:  a and b
t = 5.0805, df = 17, p-value = 4.632e-05
alternative hypothesis: true difference in means is greater than 27
95 percent confidence interval:
 57.24907      Inf
sample estimates:
mean of the differences
                     73

>  t.test( a,b,m=54.7,alternative ="greater",paired=TRUE,var.equal=TRUE)

        Paired t-test

data:  a and b
t = 2.0211, df = 17, p-value = 0.02965
alternative hypothesis: true difference in means is greater than 54.7
95 percent confidence interval:
 57.24907      Inf
sample estimates:
mean of the differences
                     73
```

Figure 4A4:

```
> ##### BASE MODEL ####
>
> mylm   = lm( y ~ 1 + B,  data=temp_data)
> summary(mylm)

Call:
lm(formula = y ~ 1 + B, data = temp_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.0381 -1.1512  0.6746  1.5003  2.8266 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 104.9804     0.3528  297.56   <2e-16 ***
B            24.3664     1.6087   15.15   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.11 on 34 degrees of freedom
Multiple R-squared:  0.8709,    Adjusted R-squared:  0.8671 
F-statistic: 229.4 on 1 and 34 DF,  p-value: < 2.2e-16

> ## Does the effect of changes in sample.size significantly help Balance to predict output?
> ## Ho: No significant effect
> ## H1: at least one significant effect
>
> ## 1. The effects of sample.size and interaction with B are significant?
>
> mylm1  = lm( y ~ 1 + B*sample.size,data=temp_data)
> anova(mylm,mylm1)
Analysis of Variance Table

Model 1: y ~ 1 + B
Model 2: y ~ 1 + B * sample.size
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     34 151.35                           
2     32 140.82  2    10.524 1.1957 0.3156
> ## 2. Only the effect of sample.size on B is significant?
>
> mylm2  = lm( y ~ 1 + B*sample.size - sample.size,data=temp_data)
> anova(mylm,mylm2)
Analysis of Variance Table

Model 1: y ~ 1 + B
Model 2: y ~ 1 + B * sample.size - sample.size
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     34 151.35                           
2     33 141.01  1    10.332 2.4179 0.1295
```

```
> ## 3. The effect of sample.size and B
>
> mylm3   = lm( y ~ 1 + B + sample.size,  data=temp_data)
> anova(mylm,mylm3)
Analysis of Variance Table

Model 1: y ~ 1 + B
Model 2: y ~ 1 + B + sample.size
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     34 151.35
2     33 149.74  1    1.6111 0.3551 0.5553
> ## 4. The effects of sample.size on period
>
> mylm4   = lm( y ~ 1 + B + period*sample.size,  data=temp_data)
> anova(mylm,mylm4)
Analysis of Variance Table

Model 1: y ~ 1 + B
Model 2: y ~ 1 + B + period * sample.size
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     34 151.35
2     31 116.84  3     34.51 3.0521 0.04306 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> mylm5   = lm( y ~ 1 + B*period*sample.size,  data=temp_data)
> anova(mylm,mylm5)
Analysis of Variance Table

Model 1: y ~ 1 + B
Model 2: y ~ 1 + B * period * sample.size
  Res.Df     RSS Df Sum of Sq     F Pr(>F)
1     34 151.347
2     28  93.601  6    57.746 2.879 0.0259 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(mylm5)

Call:
lm(formula = y ~ 1 + B * period * sample.size, data = temp_data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2352 -0.6293  0.1445  0.7560  3.7653

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           106.679036   9.519699  11.206 7.36e-12 ***
B                      27.651147  82.099146   0.337    0.739
period1                 6.268438  12.739403   0.492    0.627
sample.size             0.002933   0.015696   0.187    0.853
B:period1              31.124178  89.884112   0.346    0.732
B:sample.size          -0.039690   0.130538  -0.304    0.763
period1:sample.size    -0.019562   0.022148  -0.883    0.385
B:period1:sample.size  -0.027555   0.146328  -0.188    0.852
---
```

```
> mylmfinal = step(lm( y ~ B*sample.size*period, data=temp_data  ),direction="backward")
Start:  AIC=50.4
y ~ B * sample.size * period

                       Df Sum of Sq    RSS    AIC
- B:sample.size:period  1   0.11854 93.720 48.444
<none>                              93.601 50.399

Step:  AIC=48.44
y ~ B + sample.size + period + B:sample.size + B:period + sample.size:period

                     Df Sum of Sq     RSS    AIC
- B:sample.size       1    3.6482  97.368 47.819
- sample.size:period  1    3.7447  97.464 47.855
<none>                             93.720 48.444
- B:period            1    8.5378 102.257 49.583

Step:  AIC=47.82
y ~ B + sample.size + period + B:period + sample.size:period

                     Df Sum of Sq     RSS    AIC
- sample.size:period  1    0.9881  98.356 46.183
<none>                             97.368 47.819
- B:period            1   19.4696 116.837 52.382

Step:  AIC=46.18
y ~ B + sample.size + period + B:period

              Df Sum of Sq     RSS    AIC
- sample.size  1    0.7556  99.111 44.458
<none>                      98.356 46.183
- B:period     1   18.6793 117.035 50.442

Step:  AIC=44.46
y ~ B + period + B:period

           Df Sum of Sq     RSS    AIC
<none>                   99.111 44.458
- B:period  1    25.371 124.482 50.663
> mylmfinal $anova
                    Step Df  Deviance Resid. Df Resid. Dev      AIC
1                        NA        NA        28   93.60109 50.39883
2 - B:sample.size:period  1 0.1185411        29   93.71963 48.44439
3        - B:sample.size  1 3.6482145        30   97.36784 47.81917
4   - sample.size:period  1 0.9880745        31   98.35592 46.18266
5          - sample.size  1 0.7555815        32   99.11150 44.45816
> summary(mylmfinal)

Call:
lm(formula = y ~ B + period + B:period, data = temp_data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4234 -0.6667  0.1100  0.9736  2.6775
```

Figure 4A5

Examining whether sample size alone or between periods has any significant effect in the nature of the forecast error:

```
> mylm3_e = lm(e~ 1 + sample.size*period,data=temp_data)
> summary(mylm3_e)

Call:
lm(formula = e ~ 1 + sample.size * period, data = temp_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.3558 -1.2469  0.1741  1.4506  2.9620 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.638411   4.585350   1.448    0.157
sample.size        -0.010317   0.007801  -1.323    0.195
period1            -4.339427   9.494305  -0.457    0.651
sample.size:period1  0.005012   0.017042   0.294    0.771

Residual standard error: 2.019 on 32 degrees of freedom
Multiple R-squared:  0.138,     Adjusted R-squared:  0.05721 
F-statistic: 1.708 on 3 and 32 DF,  p-value: 0.185

> step(mylm3_e, direction="backward")
Start:  AIC=54.35
e ~ 1 + sample.size * period

                     Df Sum of Sq    RSS    AIC
- sample.size:period  1   0.35264 130.81 52.448
<none>                            130.46 54.351

Step:  AIC=52.45
e ~ sample.size + period

              Df Sum of Sq    RSS    AIC
- sample.size  1    7.2778 138.09 52.397
<none>                     130.81 52.448
- period       1   19.0242 149.83 55.336

Step:  AIC=52.4
e ~ period

         Df Sum of Sq    RSS    AIC
<none>                138.09 52.397
- period  1    13.259 151.35 53.698

Call:
lm(formula = e ~ period, data = temp_data)

Coefficients:
(Intercept)      period1  
     0.6069      -1.2138  
```

We re-run the procedure without the intercept

```
> mylm3_e = lm(e~ -1 + sample.size*period,data=temp_data)
> summary(mylm3_e)

Call:
lm(formula = e ~ -1 + sample.size * period, data = temp_data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3558 -1.2469  0.1741  1.4506  2.9620

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
sample.size         -0.010317   0.007801  -1.323    0.195
period0              6.638411   4.585350   1.448    0.157
period1              2.298984   8.313627   0.277    0.784
sample.size:period1  0.005012   0.017042   0.294    0.771

Residual standard error: 2.019 on 32 degrees of freedom
Multiple R-squared:  0.138,     Adjusted R-squared:  0.03028
F-statistic: 1.281 on 4 and 32 DF,  p-value: 0.298


> step(mylm3_e, direction="backward")
Start:  AIC=54.35
e ~ -1 + sample.size * period

                     Df Sum of Sq    RSS    AIC
- sample.size:period  1   0.35264 130.81 52.448
<none>                            130.46 54.351

Step:  AIC=52.45
e ~ sample.size + period - 1

              Df Sum of Sq    RSS    AIC
- sample.size  1    7.2778 138.09 52.397
<none>                     130.81 52.448
- period       2   20.5244 151.33 53.695

Step:  AIC=52.4
e ~ period - 1

         Df Sum of Sq    RSS    AIC
- period  2    13.259 151.35 51.698
<none>                138.09 52.397

Step:  AIC=51.7
e ~ 1 - 1


Call:
lm(formula = e ~ 1 - 1, data = temp_data)

No coefficients
```

```
> mylm3_e = lm(e~ -1 + sample.size*period,data=temp_data)
> summary(mylm3_e)

Call:
lm(formula = e ~ -1 + sample.size * period, data = temp_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.3558 -1.2469  0.1741  1.4506  2.9620 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
sample.size         -0.010317   0.007801  -1.323    0.195
period0              6.638411   4.585350   1.448    0.157
period1              2.298984   8.313627   0.277    0.784
sample.size:period1  0.005012   0.017042   0.294    0.771

Residual standard error: 2.019 on 32 degrees of freedom
Multiple R-squared:  0.138,	Adjusted R-squared:  0.03028 
F-statistic: 1.281 on 4 and 32 DF,  p-value: 0.298
```

```
> step(mylm3_e, direction="backward")
Start:  AIC=54.35
e ~ -1 + sample.size * period

                     Df Sum of Sq    RSS    AIC
- sample.size:period  1   0.35264 130.81 52.448
<none>                            130.46 54.351

Step:  AIC=52.45
e ~ sample.size + period - 1

              Df Sum of Sq    RSS    AIC
- sample.size  1    7.2778 138.09 52.397
<none>                     130.81 52.448
- period       2   20.5244 151.33 53.695

Step:  AIC=52.4
e ~ period - 1

         Df Sum of Sq    RSS    AIC
- period  2    13.259 151.35 51.698
<none>                138.09 52.397

Step:  AIC=51.7
e ~ 1 - 1


Call:
lm(formula = e ~ 1 - 1, data = temp_data)

No coefficients
```

The results from the above procedures Figures 4A2, 4A3, 4A4 and 4A5 support the claim that the changes in ITS sample size do not affect the forecasting performance of survey data significantly. Furthermore, the "period" (normal vs crisis) seems to have an effect on the performance of survey data.

The backward elimination process starts with the full unrestricted model and in each step takes out one effect and re-examines if that effect increased the performance of the remainder model based on the AIC criterion. The procedure stops if : while excluding an effect the performance of the remainder model is worse. The stepwise procedure was also implemented resulting in a similar results without indicating sample size as significant effect. Also a regression of the forecasting error against the effects of change in the sample size was also considered and the backward elimination model ends up with just the intercept.

Appendix B

```
>
> Table B: Example of the Dataset in R
> head(data)
  Year Quarter     y    Re    Fe    Rp    Fp
1 1991      Q1 -3.80 0.182 0.263 0.090 0.520
2 1991      Q2 -6.10 0.202 0.293 0.101 0.465
3 1991      Q3 -6.10 0.202 0.182 0.101 0.384
4 1991      Q4 -4.10 0.172 0.212 0.111 0.404
5 1992      Q1 -1.90 0.232 0.182 0.184 0.276
6 1992      Q2  0.01 0.190 0.220 0.178 0.317
> .         .     .     .     .     .     .
> .         .     .     .     .     .     .
> .         .     .     .     .     .     .
> data[out_of_sample_period,]
   Year Quarter     y     Re    Fe     Rp    Fp
70 2008      Q2  -1.5 0.2079 0.277 0.2800 0.290
71 2008      Q3  -2.7 0.1400 0.450 0.1600 0.450
72 2008      Q4  -7.2 0.0707 0.505 0.1600 0.490
73 2009      Q1 -12.4 0.1212 0.444 0.0606 0.596
74 2009      Q2 -10.8 0.1818 0.323 0.1200 0.430
```

Table B shows a visual example on the dataset that will be used for the quantification procedures. It is important to mention some symbolism differences that are observed in the dataset. The output growth is symbolised as y instead of x and is measured in % growth from quarter to quarter. The prospective series are symbolized as R^e, F^e represent the % of firms expectations of a "*Rise*" ("*Fall*") of y which is equivalent to an "*Up*" ("*Down*") movement of over the next quarter as ticked in the ITS survey. The retrospective series are symbolized as

R^p, F^p instead of are symbolized as R , F in the report. An example is outlined below to help the reader understand the structure of the dataset in **R** and the connection with the report.

Example: In the line 1 of the dataset is 1991Q1

y(1991 Q1) : -3.8 % growth was observed from 1991Q1

Re(1991 Q1): 20.79% of the firms in 1991Q1 expect y to *"Rise"* in 1991Q2

Fe(1991 Q1): 27.7% of the firms in 1991Q1 expect y to *"Fall"* in 1991Q2

Rp(1991 Q1): 9% of the firms in 1991Q1 reported that y has "*risen*" in 1991Q1

Fp(1991 Q1): 52% of the firms in 1991Q1 reported that y has "*fallen*" in 1991Q1

Table B1: *Descriptive Statistics Analysis*							
Whole Sample	Prospective data [1991Q1 : 2009Q2]			Retrospective data [1991Q1 : 2009Q2]			Official Data [1991Q1 :2009Q2]
Nobs = 74	R^e	S^e	F^e	R^p	S^p	F^p	x_t
mean	24.8	54	20.6	23.5	48	27	0.005
median	25	54	19	24	50	25	0.9
sd	6.1	4.6	8.1	7.2	5	9.7	3.1
Min	7	41	9	6	34	9	-12.4
max	38	63	50	42	63	59	6.7
skew	-0.44	-0.50	1.5	-0.22	-0.63	0.89	-1.5
kurtosis	-0.13	0.48	2.5	-0.14	1.06	0.66	3.7
In sample	Prospective data [1991Q1 : 2007Q4]			Retrospective data [1992Q2 : 2008Q1]			Official Data [1992Q2 :2008Q1]
Nobs = 68	R^e	S^e	F^e	R^p	S^p	F^p	x_t
mean	25.6	54.6	19.1	24.	49.4	26	0.57
median	26	55	18	24	50	24	1
sd	5.5	4	6	7	4.3	8.4	2.2
min	12	43	9	9	38	9	-6.1
max	38	63	43	42	63	52	6.7
skew	-0.27	-0.21	1.1	-0.2	-0.25	0.68	-0.52

kurtosis	-0.47	0.04	2.01	-0.045	1.07	0.32	1.27
Out of sample	Prospective data [2008Q1 : 2009Q1]			Retrospective data [2008Q1 : 2009Q1]			Official Data [2008Q2 :2009Q2]
Nobs = 5	R^e	S^e	F^e	R^p	S^p	F^p	x_t
mean	15.4	46.2	38	17.8	40.4	41.6	-6.92
median	14	43	44	16	39	45	-7.2
sd	6.50	5.8	11.7	8.3	7.1	13.3	4.8
min	7	41	23	6	34	26	-12.4
max	23	53	50	28	51	59	-1.5
skew	-0.014	0.26	-0.25	-0.15	0.47	-0.006	0.02
kurtosis	-1.99	-2.22	-2.1	-1.70	-1.66	-2.03	-2.13

In this table it clearly observed the impact on crisis on output when crisis period is omitted the descriptive statistics show noticeable differences. Also some key features are highlighted. The distribution of x has no evidence of being normal in any sample variation. Normal distribution has skewness 0, kurtosis 3 and the mean equals the median. In the whole sample period the kurtosis is close but the large negative skewness and the difference between the mean and median shows the asymmetry caused by the large negative values from the crisis is still substantial. In the *in sample* period the output is gathered around zero and is way more peaked than a normal distribution.

```
> Table B2: Unresticted Anderson Model AR(1)
> summary(UAM_AR)
Generalized least squares fit by maximum likelihood
  Model: y ~ -1 + R + F
  Data: temp
       AIC      BIC    logLik
  224.3424 233.2204 -108.1712

Correlation Structure: AR(1)
 Formula: ~1
 Parameter estimate(s):
      Phi
0.7525342

Coefficients:
      Value Std.Error   t-value p-value
R  6.130689  1.935680  3.167201  0.0023
F -5.373670  1.815417 -2.960020  0.0043

 Correlation:
  R
F -0.357

Standardized residuals:
       Min         Q1         Med          Q3         Max
-2.5984438 -0.2551517  0.3748480  0.7964871  2.5599281

Residual standard error: 1.792088
Degrees of freedom: 68 total; 66 residual
```

Table B2 is a summary of the output from the UAM regression (13) A2 in **R**[24]. In fact the output comes from a Generalised Least Squares estimation using the gls() function in **R**. The parameters $\hat{a} = 6.130$, $\mathrm{se}(\widehat{\alpha}) = 1.9$ and $\hat{b} = -5.373, se(\hat{b}) = 1.815$ are adjusted for autocorrelation and Phi denotes the $u_t \sim AR(1)$ parameter $\hat{\varphi} = 0.752$ for the error term. Both variables (R_t, F_t) are found to be significant $p - value = 0.0023$, $p - value = 0.0043$.

[24] **R** is a open source statistical program.

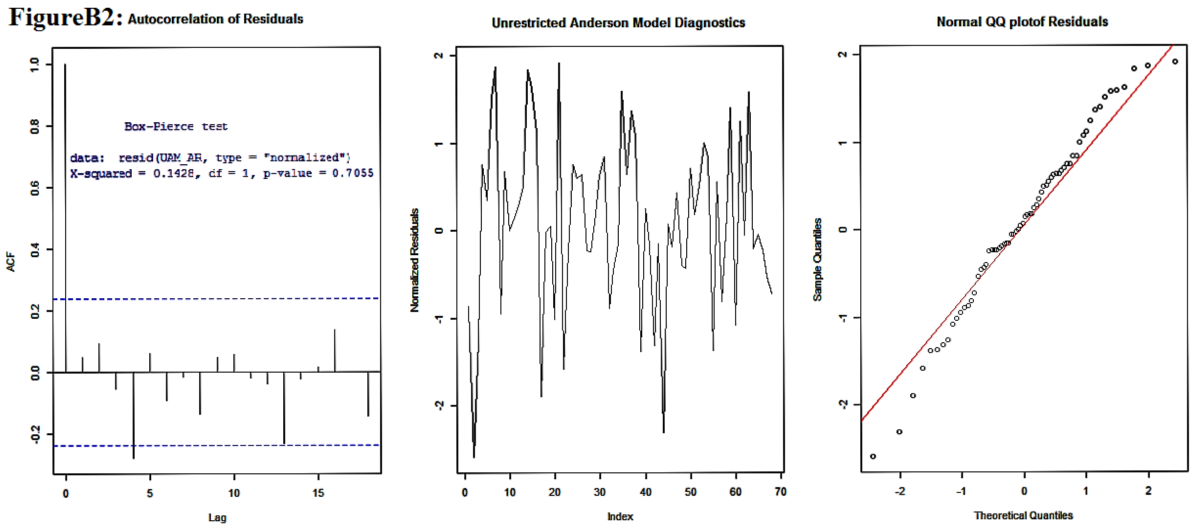


Figure B2 shows the diagnosstic test for the UAM model. It is evident throught the acf and the Box-Pierce p.value that there is no evidence to reject (H_0) the null hypothesis of autocorrelation in the residuals. Following by the two plots on the right the homoskdasticity and normality assumptions are not clear mainly because of the large negative values of that caused by the early 1990’s recession.