Field of the paper ABC, a black cat; DEF, doesn't ever fret; GHI, goes home immediately. Author One PhD, Department, Institution, City, State or Province, Postal Code, Country correspondingauthor@email.com Department, Institution, City, State or Province, Postal Code, Country Funder One, Funder One Department, Grant/Award Number: 123456, 123457 and 123458; Funder Two, Funder Two Department, Grant/Award Number: 123459

Repeated measurements with unintended feedback

Richard D. Gill[1]

[1]Affiliation not available

April 19, 2021

Abstract

Abstract. An econometric analysis of consumer research data which hit newspaper headlines in the Netherlands illustrates almost everything that can go wrong when standard statistical models are fit to the superficial characteristics of a data-set with no attention paid to the data generation mechanism.

Author: Richard D. Gill, Mathematical Institute, Leiden University, The Netherlands; email address: gill@math.leidenuniv.nl.

Dedication. This paper is dedicated to my collaborator and friend of many years, Ørnulf Borgan (University of Oslo), on the occasion of his virtual 65th birthday celebrations.

Funding information. Some of this research was commissioned and funded by law company Bird and Bird, Amsterdam.

Keywords — CAUSALITY VERSUS CORRELATION, DATA GENERATION MECHANISM, FEEDBACK IN SAMPLING AND MEASUREMENT, PREDITOR-PREY CYCLES, QUESTIONABLE RESEARCH PRACTICES, UNHEALTHY RESEARCH STIMULI

Introduction

Lifetime data is often collected over a long period of calendar time. As time goes by, data-gathering procedures may change ... and they may change as a response to continuous data monitoring. How can one tease the different effects apart?

If one is only interested in *describing* the observed past, maybe it doesn't matter. Statistical analysis can reveal *parsimonious descriptions* of past data. But if politicians or other agents use the results to push for policy interventions, it may matter a great deal.

In this paper I will discuss an extreme example from micro-economics, in which an annual survey was carried out and in which the "units" (small businesses: fishmongers and supermarkets) sampled in any year, and moreover *how* they were evaluated, were possibly strongly influenced by the analysis results obtained the previous year. The probability of being sampled varies from year to year in a way which depends on what has happened in the past. To complicate matters further, the way in which the units are evaluated (by a tasting panel) presumably can change over the years. Moreover there is a life-time survival aspect – some small businesses fail, and new ones are started, in response to annual publication of the perceived quality of their products. This can lead to phenomena reminiscent of the preditor-prey population cycles observed in ecology: the interacting populations of snowshoe hare and Canadian lynx providing the paradigmatic example. More examples come from quantum physics where measuring a system disturbs it so fundamentally that the question arises, and is hotly debated to this day, does the system have any intrinsic properties at all?

The product in question is “Dutch New Herring”, and the main sources (unfortunately, in Dutch) are two discussion papers published by Tilburg University, and two items in a data repository, Vollaard (2017a, 2017b, 2020). I will later go into more depth into what is meant by that three word phrase, and I capitalise the words in order to emphasize that it is the legally protected name in the EU of a commercial product. For the moment it suffices to say the following. Every nation around the North sea has traditional ways of preparing North Atlantic herring. For centuries, herring has been a staple diet of the masses. It is typically caught when the North Atlantic herring population comes together at its spawning grounds, one of them being in the Skagerak, between Norway and Denmark. Just once a year there is an opportunity for fishers to catch enormous quantities of a particular extremely nutricious fish. The fishers have to preserve their catch during a long journey back to their home base; and if the fish is going to be consumed by poor people throughout a long year, further means of conservation are required. Dutch, Danish, Norwegian, British and German herring fleets (and more) all compete for the same fish; but what people in those countries eat varies from country to country. Traditional local methods of bringing ordinary food to the tables of ordinary folk become cultural icons, tourist attractions, gastronomic specialities, and export products.

The experience of statisticians working in survival analysis has taught us how important it is to model the data generating process as something on top of statistical modelling of the underlying system of interest. Even so, the medical literature is full of routine applications of the Cox regression model, processed with the help of standard statistical packages, and in which no thought at all has been given to the modelling of the data generation. The choice of analysis has come to be made on the basis of the formal structure of the data-base. Times of events of interest, censoring indicators, and covariates ... press the button and publish the results.

One can see something similar happening in many scientific fields. In micro-economics one collects data on a “sample” of firms; there is a dependent variable (the variable of main economic interest – the variable to be “explained”), and a whole lot more “explanatory variables” or covariates. The analyst thinks “regression analysis”; and the choice of regression model – led by multiple choice questions put by the software package – will depend on formal properties of the data. Is the “dependent variable” binary, categorical, or continuous?

There is a major conflict here between *prediction* and *understanding* or *explanation* . If one merely wants to successfully predict outcomes of future observations perhaps it doesn’t matter. But if one wants to predict what would be the effect of an intervention, we enter the field of causality. This also applies to counter-factual interventions which could conceivably have been made in the past, but in actual fact weren’t. The task of law courts, both criminal and civil, is to determine what would have happened if certain actors had performed different acts. In general, this requires understanding of causal mechanisms. Such understanding can be gained from statistical modelling but it is hardly possible without being supported by prior theoretical understanding. For instance, at the very least, we tend to believe that cause and effect works forwards in time. We tend to forget that sampling from end-results can reverse the apparent direction of causation. Statisticians shield themselves from responsibility by claiming that they can only determine correlation, not causation. But their clients are only interested in causation. A consulting statistician learns instinctively how to please clients. The boiler-plate small print leaves the statistician free from responsibility for what is done with the correlations which are discovered. They will be causally interpreted.

A second theme of this paper is the topic of scientific integrity and of questionable research practices. Science publication is obviously a central part of academic life but it is driven to a large extent by the necessity to maintain the infrastructure which makes it possible: funding. In order to do research you will need to find someone who is prepared to pay you to do it.

The first sections of this paper will sketch further background of a particular case in which the author was involved as consultant to a law firm. The law firm was acting for a national newspaper, and the newspaper was fighting an individual university academic, an economist, who had successfully created a big media stir by reporting his statistical analysis of the data gathered by the newspaper to annually rank Dutch New Herring sales outlets. The result was that the economist appeared on current affairs talk shows, and the newspaper suspended its annual evaluation and suffered damage to its reputation and circulation. The lawyers were able to trigger investigations of possible violations of scientific integrity first at university level and then at national level, but they did not result in “conviction”. They did finish with the advice to carry out further research. So far, this has not happened. In my opinion, there is an enormous amount to be learnt from this case both about analysis of causality and about scientific ethics, and in particular about "perverse scientific stimuli" by which I think of the pressure on academics to produce results which create media publicity for their institution. In my opinion, the pendulum has swung so far towards the notion that scientific research must be justified by immediate public appeal and rapid social impact, that current research practices are harming science, scientists, and society.

Vollaard's analyses

Traditionally, the Dutch herring fleet brings in the first of the new herring catch mid June. The very first catch is auctioned and a huge price (given to charity) is paid for the very first barrel. Very soon, fishmongers, from big companies with a chain of stores and restaurants to small businesses selling fish in street markets are offering Dutch New Herring to their customers. It's a traditional delicacy. For a number of years, a Rotterdam based newspaper *Algemene Dagblad* (referred to as AD in the sequel) has been carrying out an annual comparison of the quality of the product offered in a sample of consumer outlets. A small team of expert herring tasters pays surprise visits to the typical small fishmonger's shops and market stalls where customers can order portions of fish and eat them on the premises (or even just standing in a busy food market). The team evaluates how well the fish has been prepared, preferring especially that the fish have not been cleaned in advance but that they are carefully and properly prepared in front of the client. They judge the taste and check the temperature at which it is given to the customer (by law it may not be above 7 degrees). A sample is sent to a lab for a number of measurements: weight, fat percentage, signs of microbiological contamination. They are also interested in the price (per gram). An important characteristic is "ripeness". The organs of the fish were removed when they were caught, and the fish kept in lightly salted water. But one internal organ was left, a fish's equivalent to our pancreas. It contains enzymes which slowly transform some of the protein into fat and this process is responsible for a special almost creamy taste which is much treasured by the Dutch consumers (and is apparently uniquely Dutch). This ripening process might have been just enough, quite a lot, too much or much too much.

This information all gets written down and combined subjectively (the team averages the scores given by its members) to produce a score from 0 to 10, where 10 is perfection; below 5.5 is a failing grade. The outlets which have taken part are ranked and the ranking is published in the newspaper. Coming out on top is like getting a Michelin star. The outlets at the bottom of the list may as well close down straight away. One sees from the histogram below that in 2016 and 2017, more than 40% of the outlets got a failing grade. The distribution looks nicely smooth except for the peak of nearly 10% of outlets which got a zero, which means that their wares did not satisfy the minimal legal health requirements.

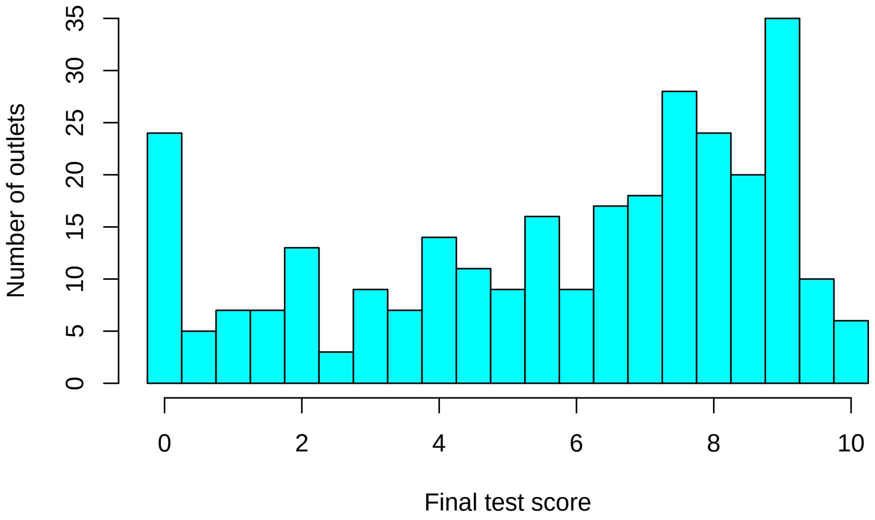

Figure 1: Histogram of final test scores 2016 and 2017, N=144+148=292 scores.

Now, in recent years, there has been more and more acrimonious criticism of the AD test. As one can imagine, it is mainly the owners of outlets who get bad scores who are unhappy about the test. Many of them, perhaps justly, are proud of their product and have many satisfied customers too. Various accusations are therefore flung around. The most serious one is that the testing team is biased and indeed has a conflict of interest. The lead taster gives courses on preparation of Dutch New Herring and indeed led the movement to have the "brand" registered with the EU. There is no doubting his expertise, but he has been hired by one particular wholesale business, owned by a successful businessman of Turkish extraction, which as one might imagine leads to jealousy and suspicion. Especially since the 10 retail outlets of fish supplied by that particular company regularly get very high grades indeed in the annual AD Herring Test. Other accusations are that the herring tasters favour business in the neighbourhood of Rotterdam (home base of the AD); as herring congnoscenti know, the people in various Dutch localities have slightly different tastes in Dutch New Herring. There is an ancient rivalry between Amsterdam and Rotterdam (it is not restricted to Ajax versus Feyenoord).

In 2017 a young Dutch econometrist from Tilburg University by the name of Ben Vollaard entered the fray. The story goes that he appreciates a decent Dutch New Herring and that his favourite fishmonger complained about the 2017 ranking. He had a student collect all the data published in the last two years by the AD, and put together a little spreadsheet of 292 observations of 21 variables (actually, some of the 21 variables are simple transformations of others). He then ran a regression analysis, with dependent variable being the final test grade, and with various characteristics of the fish served in each of those outlets as explanatory variables. Some of these variables are pretty objective measurements (temperature at which the fish was served, measurement of microbiological contamination (presence of harmful bacteria), price per hundred gram, weight per portion (i.e., per fish), fat percentage. Others are variables subjectively allocated by the fish tasters such as the degree to which the product has matured, and how well the fish has been cleaned. Also they note whether they could observe the fish being cleaned on the premises as each client orders them (which is how tradition dictates it should be done).

The data actually comes from the tests of two years and many of the sales outlets participate in the test year after year. Thus one can expect that most of the observations come in pairs, and that within each pair there is a high similarity of all the measurement outcomes.

But this did not deter Mr. Vollaard. He prepared a report based on the results of a single multiple regression analysis and proceeded to draw attention to it in the media, encouraged by his university (which put out extremely tendentious and attention grabbing press releases). A few months after the first report, he did another multiple regression analysis, and again proceeded to get attention in social media. This led to appearances on Dutch daily current affairs programs and to attention even from foreign media such as a big spread in *The Economist*. Dr. Vollaard repeatedly stated to journalists and interviewers that he was only looking at correlations and his methodology did not allow one to draw causal inferences from them but (a) the testing team had a conflict of interest and (b) he thought that the AD Herring Test stinks. In other words, as a scientist who had performed a sophisticated statistical analysis he wasn't going to say out loud that his results showed that the test team were biased and that their bias influenced their ranking, but he certainly believed that himself, and he saw a heap of evidence for that in his statistical modelling of the actual data.

In my opinion this behaviour does violate scientific integrity, though some of the blame must go to the university's PR department's press releases. Moreover, as I will now go on to explain, I think that his inferences from his regression model were unwarranted and that the analyses were of such questionable value as to make them utterly worthless. What should have happened, but never happened till long, long, later, was to publish his data. His reports appeared in a series of working papers of his university, they never received peer review, let alone got published in a scientific journal.

In the meantime, under the deluge of negative publicity, the AD announced that they would now stop their annual herring test. They did hire a law company to try to bring an accusation of failure of scientific integrity to Tilburg University's "Commission for Scientific Integriity". The law firm approached me for advice. I was initially extremely hesitant to be a hired gun in an attack on a fellow academic but as I got to understand the data and the analyses and the subject matter, I had to agree that the AD had a point. Moreover, various aggrieved herring sellers were following up with their own civil action against the AD; and the sales outlet which did so well in the test, also started a civil action against Tilburg University, since its own reputation was damaged by the whole affair. It was quite a storm in a small barrel of old herrings.

To my amazement, no other Dutch statistician or econometrician got involved in the case at all. I think I found this the most disturbing thing of all. I gave talks about the case at a number of seminars, and also approached by own university's PR department to get some advice and even training on how a scientist should enter a societal fight.

Here is the main result of Vollaard's first report, nicely reproduced by "R".

```
lm(formula = finalscore ~
                    weight + temp + fat + fresh + micro +
                    ripeness + cleaning + yr2017)

Residuals:
     Min      1Q  Median      3Q     Max
 -4.0611 -0.5993  0.0552  0.8095  3.9866

Residual standard error: 1.282 on 274 degrees of freedom
Multiple R-squared:  0.8268, Adjusted R-squared:  0.816
F-statistic: 76.92 on 17 and 274 DF,  p-value: < 2.2e-16

```

```
Coefficients:
                 Estimate   Std.Error  t value Pr(>|t|)

Intercept         4.139005   0.727812   5.687 3.31e-08 ***

weight (grams)    0.039137   0.009726   4.024 7.41e-05 ***

temp
    < 7 deg         reference-category
    7 -- 10      -0.685962   0.193448  -3.546 0.000460 ***
    > 10 deg     -1.793139   0.223113  -8.037 2.77e-14 ***

fat
    < 10            reference-category
    10--14        0.172845   0.197387   0.876 0.381978
    > 14          0.581602   0.250033   2.326 0.020743 *

fresh             1.817081   0.200335   9.070  < 2e-16 ***

micro
    very good       reference-category
    adequate     -0.161412   0.315593  -0.511 0.609443
    bad          -0.618397   0.448309  -1.379 0.168897
    warning      -0.151143   0.291129  -0.519 0.604067
    reject       -2.279099   0.683553  -3.334 0.000973 ***

ripeness
    mild            reference-category
    average      -0.377860   0.336139  -1.124 0.261947
    strong       -1.930692   0.386549  -4.995 1.05e-06 ***
    rotten       -4.598752   0.503490  -9.134  < 2e-16 ***

cleaning
    very good       Mathematreference-category
    good         -0.983911   0.210504  -4.674 4.64e-06 ***
    poor         -1.716668   0.223459  -7.682 2.79e-13 ***
    bad          -2.761112   0.439442  -6.283 1.30e-09 ***

yr2017            0.208296   0.174740   1.192 0.234279
--
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

No surprises here. The testing team prefers fatty and larger herring, properly cooled, mildly matured, freshly prepared and well cleaned. We have a delightful amount of statistical significance.

I will add to the estimated regression model also the standard plots. Mr. Vollaard apparently did not carry out any model checking.

There are some serious statistical issues. There seem to be a couple of serious outliers. But we also know that the observations come almost all in pairs – the same outlet evaluated in two subsequent years. The data set has been anonimized too much. Each outlet should have been given a random code so that one can identify the pairs and take account of possible dependence from one year to the next.

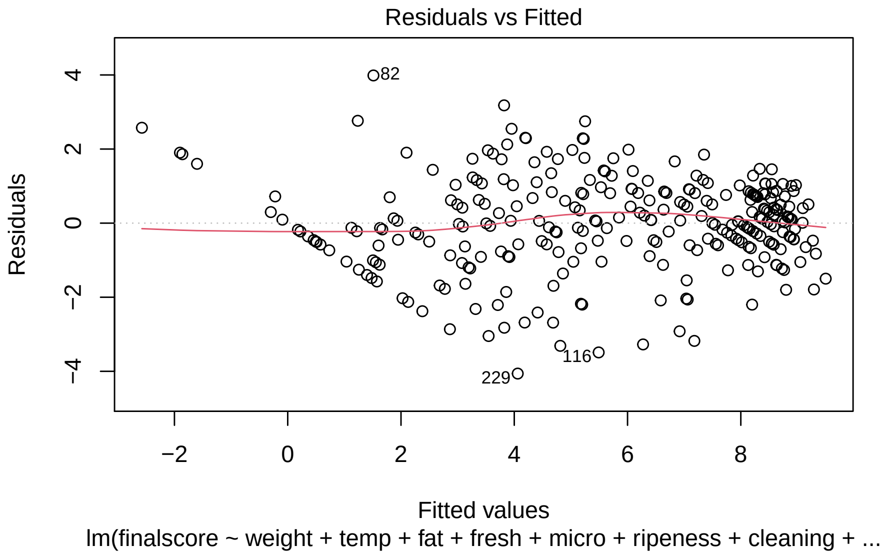

Figure 2: Residuals vs fitted values

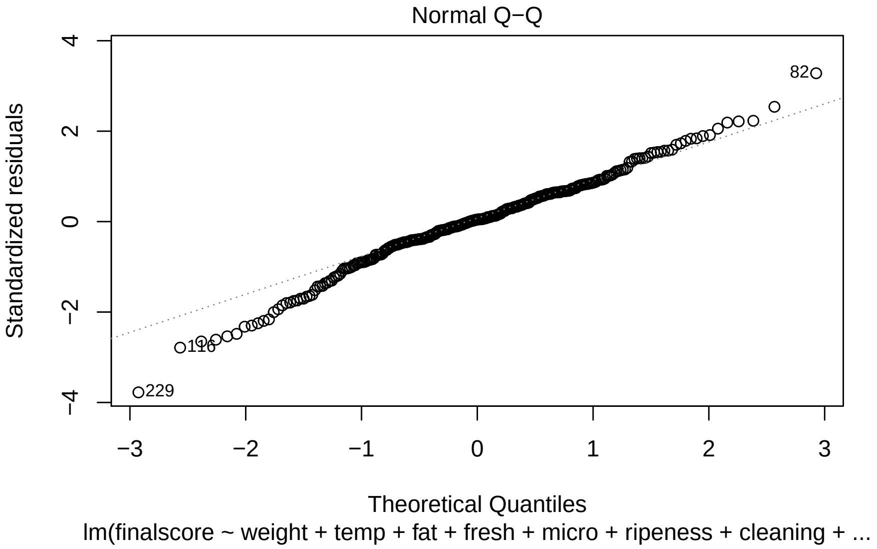

Figure 3: QQ plot (standardized residuals versus standard normal quantiles)

There is a serious issue with the observations which got a final score of zero. One could better say that those outlets were disqualified on grounds of violation of basic hygiene laws. The model should have been split into two parts: a linear regression model for the scores of the not-disqualified outlets; a logistic regression model, perhaps, for predicting "disqualification". But this seems to be quite a waste of time. But at least it is possible to analyse each of the years separately, and to remove the "disqualified" outlets. That is easy to do. Analysing just the 2017 data, the analysis results look a whole lot cleaner; the two bad outliers have gone. I will not present the results here. The data set, now as a `.csv` spreadsheet, can be obtained from me.

But why did Mr. Vollaard come to his strong disapproval of the testing team from this data-analysis? He added a dummy variable to indicate outlets more than 30 Km from Rotterdam. It had a significant, negative coefficient. This was sufficient for him to accuse the testing team of bias towards Rotterdam outlets. I would say "so what?" The

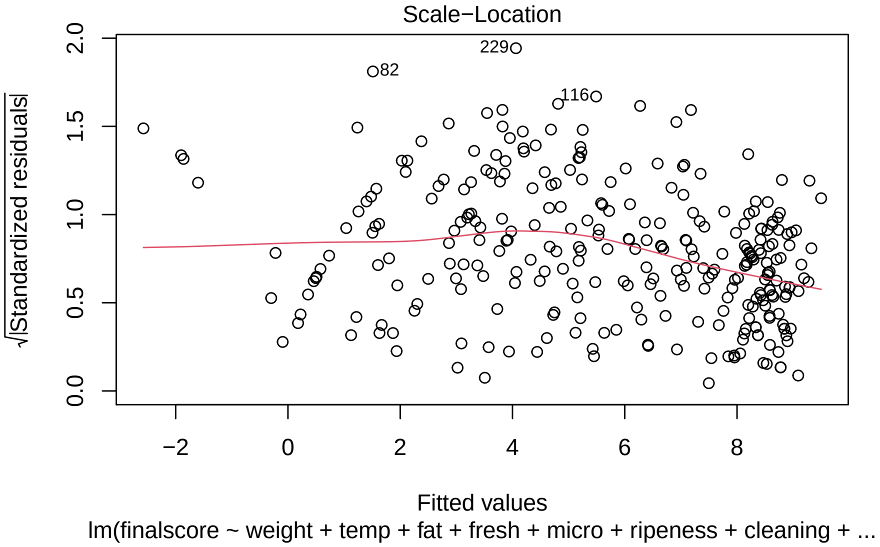

Figure 4: Square roots of standardized residuals against fitted values

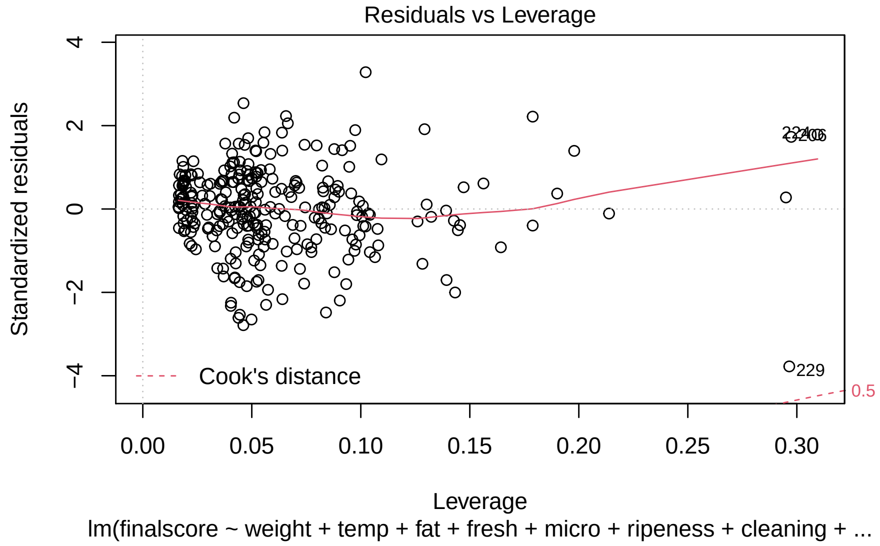

Figure 5: Standardized residuals versus leverage

herring tasters judge Dutch New Herring according to the traditional standards of the region in which their newspaper is based, and where most of their readers reside. It includes the main herring port of Scheveningen (just outside The Hague). But I will also give another reason why this result should be taken with a pinch of salt, in the next section.

He then went on a further hunt for evidence of bias. In a second report, he added a dummy variable for the 10 retailers who were clients of the wholesale company *Atlantic*; the company which had a connection with the senior herring taster. It was not statistically significant! Many of those outlets did very well and the regression model, thought of as showing us summary statistics (correlations) shows us why. They scored well on the criteria which interest the tasters.

By the way, we know that one of the *Atlantic* outlets was incorrectly classified as non-Atlantic and that in that year it had got a very bad score. It would be nice to know which observation that is.

Vollard had to think of something else in order to support his accusation of a specific bias in favour of *Atlantic* outlets. He came up with a standard econometricians' recipe for measuring the amount of variation in final score which can be attributed to different groups of explanatory variables. As I have made clear, some of the variables are results of laboratory measurements, some are the "subjective" evaluation of the three man testing and tasting team. He found that the subjectively evaluated "ripeness" or "maturation" was very important, while the objective "microbiological test" had almost no contribution to make. Also the subjectively measured "cleaning" was very important. In short, those two subjective variables took account of half of the observed variation; two objective variables (weight and fat content) took account of the other half, other objective measures were unimportant. Because the final score is, for 50%, explained by subjectively evaluated criteria, he considered the test worthless and suggested bias of the test team. In particular he put his finger on the fact that the subjectively measured ripeness had a huge effect while the objectively measured microbiological test had almost none, though to him, both are measuring the same thing: the degree to which the fish is "going off".

Now, the maturation of Dutch New Herring is a chemical process associated with the work of the enzymes from the pancreas of the fish, as well as autonomous chemical ageing. "Matured" venison, hare, wild boar meat, is preferred to fresh. Cheese is preferred when it has ripened. Whisky is preferred after many years maturation. Though the Dutch New Herring is kept at low temperature and in salty water, cell walls are slowly breaking down, various substances are diffusing through the body of the fish. Provided this process is not allowed to continue too long, it leads to changes in flavour which some consumers like, others dislike. Consumers of Dutch New Herring have different "tastes" regarding ripeness. Only if the ripening has continued for much too long can one say that the fish has gone rotten.

Also with time, the fish gets saltier, and too salty (though how much is too much is a matter of taste) is not nice either.

The microbiological measurement on the other hand tells us whether the fish has got contaminated with bacteria through e.g., contamination or careless removal of intestines, etc. This "objective" microbiological measurement tells us whether or not the fish is safe to eat. It has almost nothing whatever to do with how it tastes, unless the contamination is very big.

Could it not simply be the case that *Atlantic* imports the best herring and treats it with the care it deserves? It is not so cheap as herring from other outlets. The Atlantic outlets are not very far from Rotterdam. I have independent evidence for this claim, and if anyone would like a recommendation from me, where the best Dutch New Herring can be eaten, I will be happy to tell them.

Conflict of interest

The author reveals that he was informed that the best Dutch New Herring his brother-in-law ever ate was at a retail outlet of *Simonis* in Leiden. That outlet got their herring from the wholesaler *Atlantic*. My informant volunteered this personal subjective taste information when he heard that I was looking at the statistics of herring taste data. He had no idea that there was a media herring war going on. I have later confirmed his impression by my own test at another *Atlantic* outlet, this time in Scheveningen. I have not consulted any other herring lovers.

More seriously, the author was paid by a well known law firm for a statistical report on Vollaard's analyses. My report, dated April 5, 2018, is in Dutch, an English translation is available at my blog, https://gill1109.com/is-the-ad-herring-test-about-more-than-the-herring/.

References

References

Vollaard, B. *Gaat de AD Haringtest om meer dan de haring?* Tilburg University (2017a). https://www.math.leidenuniv.nl/~gill/haringtest_vollaard.pdf

Vollaard, B. *Gaat de AD Haringtest om meer dan de haring? Een update.* Tilburg University (2017b). https://www.math.leidenuniv.nl/~gill/haringtest_vollaard_def.pdf

Vollaard, B. Tilburg University (2020). code$_{h}aringtest.do(4KB)$, scores$_{h}aringtest_{2}016_{2}017.dta(287KB)$