A very Dutch scandal

Fengnan Gao and **Richard D. Gill** gut and de-bone the questionable statistics that destroyed a long-running culinary tradition

PicturePartners/iStockPhoto.com

The Dutch love their herring. For over 30 years, an annual herring competition organised by the newspaper *Algemeen Dagblad* (AD) was a Dutch cultural icon and led to improved standards and high sales.

However, the fortunes of the much celebrated tradition took a sudden turn on 27 November 2017, when years of bickering among stakeholders culminated in a shocking article published in the magazine *The Economist*.[1]

Though the article was brief, its title alone – "Netherlands fishmongers accuse herring-tasters of erring" – struck the final blow to the now infamous AD herring test. By this point, the public had largely accepted claims that the herring test was biased, and that the leading taster on the testing panel had acted dishonestly to promote a company called Atlantic. The test was discontinued, the leading taster retired in disgrace, and the senior responsible editor died without seeing his name cleared. The whole tragedy started with a simple regression analysis by an economist. But how and why?

Dutch new herring and the AD test

Being abundant and nutritious, herring has since long ago been a staple food for mass consumption, especially in the countries of northern Europe, including the Netherlands, and each country has its unique custom around the fish as a food item. As the legend goes, in the fourteenth century a Dutch fisherman called Willem Beukelszoon is said to have discovered *kaken* — a traditional Dutch method of handling herring. It is a process of gutting and de-boning herring that leaves two internal organs intact. An enzyme called trypsin, emitted by the remaining organs, is responsible for the "ripening" and is essential for flavour. The herring is then thrown into brine and basically pickled in its own juices for about 5 days, often in oak barrels.

In today's Netherlands, Dutch new herring is celebrated as a culinary treasure, and the first catch of the season is a highly anticipated annual event – something like the tradition of treasuring the first batch of rice from the harvest every year in parts of Asia. The herring is most appreciated when it is cleaned on the spot in front of the customer, and eaten raw whole or sliced, often with onion. The taste is salty, buttery, and slightly sweet, with a creamy texture. Being a legal designation protected by the European Union, "Dutch new herring" refers to herring that meets certain predefined criteria, in accordance with the traditional Dutch way. Moreover, herring may no longer be labelled "new" after a few months.

In the Dutch new herring season, for 36 years, the newspaper AD appointed a three-person team – the same people every year – consisting of a seasoned herring expert, a senior editor and a young journalist. Unannounced, the team would visit small fishmonger shops and market stalls where customers could order and eat portions of fish on-site. Most of the participating outlets volunteered for the test or were nominated by their customers, and vendors that were in last year's top ten list were encouraged to stay in the test.

The team assessed the preparation of the fish, with a preference for fish that were carefully cleaned and properly prepared right in front of the client. The team also evaluated the taste of the fish and checked to ensure that it was not served dangerously warm. In addition, the team sent a sample of the fish to a laboratory for several measurements, including weight, fat percentage, and signs of microbiological contamination. One key, albeit somewhat subjective, factor in the assessment was the "ripeness" of the herring. A rating on each sub-category of interest was decided collectively by the team and written down, and a *provisional* score was obtained by averaging the scores given separately by the three members to generate a rating on a scale from 0 to 10 – with 10 indicating perfection and 6 considered a pass. Outlets that failed the basic hygiene regulations (e.g., dangerous microbiological contamination) or that were too disgusting to taste received a score of 0. The participating outlets were then ranked, and the top ten ranking outlets were revisited, with provisional scores adjusted accordingly. The final scores and ranking were published in AD and made available in full online. Receiving a high ranking brought fame and more customers, while those ending up at the bottom of the list might as well have shut down.

The downfall

The episode began in 2017 with Dr Ben Vollaard, a herring enthusiast and a young economist from Tilburg University in the South of the Netherlands. After decades in operation, the AD test had gathered its fair share of opponents, typically among those who did not do well in the test but were confident about their herring. Rumours began circulating that the tasting panel was biased. This interested Dr Vollaard, who first heard of such a rumour from the local herring vendors he frequented. He started investigating these rumours and put two working papers on his university webpage in July and November 2017, respectively.[2,3] In the first paper he claimed that the testing panel was positively discriminating the outlets in the Rotterdam area, where AD is based, with the suggestion that the newspaper was manipulating the test in favour of the outlets in its home city. In the second paper, even more damaging to both the herring test and the panellists, he claimed that the panel, particularly the leading panellist Aad Taal, was promoting the high-end fish wholesale company Atlantic. Taal had a known conflict of interest due to consulting work for Atlantic, which bred speculations of dishonesty on his part. But suggesting scientific evidence of dishonesty is another matter.

Of course, Vollaard is welcome to his opinion, and a discussion paper on a personal website is not a formal scientific publication, but rather an invitation for discussion. But in this case, the public relations (PR) department in Tilburg University caught wind of Vollaard's reports and realised that a scandal was brewing with great PR potential. It put out a press release both times, exaggerating Vollaard's findings. Soon the AD herring test was in the spotlight of every Dutch news outlet and became a national dinner conversation topic.

Vollaard appeared on current affairs talk shows on national television. He became a Robin Hood hero to herring vendors who did poorly on the test. Facing mounting criticism, AD terminated the herring test, admitting that matters of taste can be a matter of taste yet staunchly denying there was any bias. The newspaper made complaints to Vollaard and to Tilburg University. In particular, they complained that his data was incorrect: one Atlantic-supplied outlet had obtained a score of 0.5, and Vollaard must have misclassified it. Vollaard refused to correct his data. Aggrieved herring outlets started lawsuits against AD.

Obviously there can be perfectly innocent explanations for public perceptions of bias. The readers of AD are mostly from Rotterdam, its base, and unsurprisingly, folks from different parts of the Netherlands have different preferences in their herring. Perhaps the tasting panel preferred flavours that are appreciated more by Rotterdammers? Yet the herring test had been running for three decades, and it should have been no secret to any participating outlets what the panel's preferences were, given that AD published all test results: the overall ranking, overall scores, scores on components, and pithy verbal jury reports on each outlet.

Vollaard's second claim – that fish from wholesalers Atlantic were being favoured by testers – could be as easily explained by the fact that Atlantic supplied their clients with herrings that in general are superior in quality, something which can be easily verified on the objective measures (see Table 1). If Atlantic-supplied outlets generally ▶

TABLE 1: Summary of Atlantic versus non-Atlantic outlets, according to AD's definition. Quantities in parentheses are the corresponding standard deviations of the variables. "Good micro." indicates that the herring is not microbiologically dangerous for human consumption, and "good temperature" indicates that the herring is served in a "good" temperature range.

Item	Atlantic ($N = 29$)	Non-Atlantic ($N = 263$)
Final score mean (std. dev.)	7.91 (2.45)	5.50 (2.96)
Fat content (%) mean (std. dev.)	12.52 (2.91)	11.81 (2.64)
Weight (g) mean (std. dev.)	66.24 (6.60)	66.44 (8.25)
Price per piece (€) mean (std. dev.)	2.02 (0.23)	2.06 (0.36)
Freshly cleaned (%)	76.9	62.0
Well cleaned (%)	86.2	70.7
Very good cleaning (%)	62.1	39.9
Good micro. (%)	100	78.7
Good temperature (%)	82.7	50.6

Fengnan Gao is an associate professor in the School of Data Science at Fudan University. Trained as a theoretical statistician, he has developed a growing interest in exploring the frontiers between statistics and its impact on societal issues. He will join the School of Mathematics and Statistics in University College Dublin in September 2023.

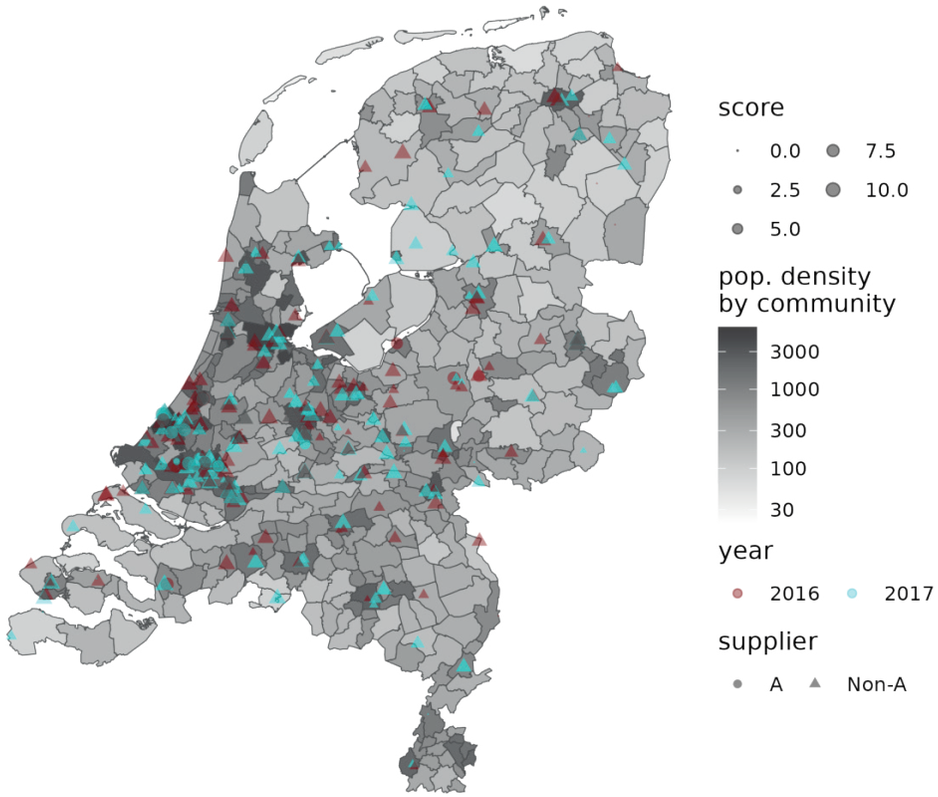

FIGURE 1: Spatial patterns of outlets and their scores. The population density of each community (*gemeente* in Dutch) is drawn in the background. Each dot (circle or triangle) represents a participating outlet, with the area of the dot reflecting its score. Supplier A refers to the outlets being supplied by Atlantic, and the Atlantic outlets are concentrated on the coast, especially in the greater Rotterdam area.

have better (more expensive?) herrings in the first place, possibly they also take care that they are better prepared and served? Moreover, being supplied by Atlantic is possibly confounded by spatial effects. Figure 1 suggests that Atlantic-supplied outlets were concentrated in coastal areas, while many outlets far from the coastal areas got dismal grades. Finally, this is not a randomised clinical trial. The herring test started as a local Rotterdam affair and due to its success slowly expanded geographically over the years. Each year there were many new contestants, and they tended to come from more distant regions; perhaps they are less familiar with the qualities which the team value most highly.

Vollaard's first analyses

AD published detailed records of past herring tests on their newspaper websites in the easily accessible HTML format. Vollaard scraped the HTML files and collected the data set consisting of all the detailed records ($N = 48 + 144$) of the herring test in both 2016 and 2017. His first report[2] conducted a simple linear regression to fit the final score to dependent variables of Vollaard's choice, including price, microbiological contamination, and fat percentage. A dependent variable in linear regression is called statistically significant if the *p*-value, that is, the probability of observing a result at least as extreme as this under the hypothesis that the variable has no effect, is below a certain threshold. Note that this probability calculation depends on many assumptions which need checking. The spotlight here was that the artificially constructed variable *k30* – a variable indicating whether the outlet is less than or more than 30 km away from Rotterdam – was statistically

This is not a randomised clinical trial. The herring test started as a local Rotterdam affair and due to its success slowly expanded geographically over the years

significant ($p \approx 0.022 < 0.05$). With this came the seemingly *natural* conclusion that the test was biased in favour of Rotterdam-based herring vendors, and public uproar followed the media attention.

The second report[3] was even more damning. Vollaard tried to access whether each outlet was supplied by Atlantic by phoning as many as possible. He made some further (unexplained) adjustments to his earlier model. With his new set-up, the Atlantic indicator was not statistically significant. But he had a new argument. He computed the expected average outcome of presumed non-Atlantic outlets, and that of presumed Atlantic outlets, according to his fitted model. Each of those two numbers is a sum of coefficients times average covariate values. The difference, about 4, is the sum of coefficients times the difference between average covariate values. He separated it into a sum over "objective" covariates and a sum over "subjective" covariates, finding roughly 2 + 2 = 4. The "subjective" variables seem responsible for half of the difference. He wrote that this gave Atlantic a two-point advantage which could *only* be the result of the panel abusing their conflict of interest.

Both Vollaard's inferences are intrinsically wrong. The first and most obvious problem is that Vollaard considers correlation as evidence of causality. Despite his repeated claim that he was only investigating correlation, he showed up on television and accused AD and the testing team of bias, which only made sense if he believed the relation was causal. Apart from this, there are many problems with a small and self-recruiting sample. There is a serious issue of possible model misspecification. Apparently, Vollaard did not conduct basic model validations. The zero grade indicated disqualification of the outlet on health regulation grounds, which depends only on a few of the variables in a deterministic

Richard Gill is an emeritus professor of mathematical statistics at Leiden University. His research interests include forensic statistics, with a focus on serial killer nurse cases, quantum information, and scientific integrity, and he is particularly passionate about rectifying injustices stemming from the unprofessional use of statistics.
Photo: Willem-Jan Schipper www.willemjanschipper.com

way. The additivity assumption that the final score is formed by adding components derived from each separate factor, and then adding uncorrelated noise, is terribly wrong. There are important complex and nonlinear dependencies which the model cannot reflect, so it fits overall very badly. We observed that the statistical significance that Vollaard relied on to make his first case is sensitive to small adjustments to his model. This issue is especially disturbing because Vollaard had chosen, in a fairly arbitrary way, discretisation of continuous variables into a few discrete categories, for instance, for the variables price and fat contents. This resulted in serious multicollinearity – that is to say, some of his artificial variables can be almost exactly predicted in a linear way from others, hence one cannot separate their linear effects on the variable of interest. Using a standard method to assess multicollinearity, we calculated the condition number of the design matrix in Vollaard[2] and the result was 31 times higher than the critical threshold above which statisticians usually consider multicollinearity to be a severe problem.[4]

Following traditional Dutch educational exam scoring, and probably in order to break ties, the AD test gave scores such as 8+ or 7– which indicate *slightly better than 8* and *slightly worse than* 7, respectively. Vollaard rounded such scores to the nearest integer (discarding information in the process). One could encode the plus/minus scores differently, such as encoding 8+ as 8.1 and 7– as 6.9. One could discretise the variables differently. All these procedures give us different significances, hitting especially the variables of most interest. Sometimes *k30* is not significant, sometimes it is; the same holds for *Atlantic*. With minimal changes to the model one can obtain many desired conclusions. All this is to be expected, because calculating statistical significances in linear regression suffers from numerical instability when the design has a large condition number.

Vollaard's second analyses

A few years later, in 2022, Vollaard, joined by a former colleague, economics professor Jan van Ours, published a paper in the *Journal of Economic Behaviour & Organization*.[5] They now claim to *prove* that the AD herring test was deliberately biased in favour of Atlantic-supplied outlets. The paper concludes that the bias of about half a point which they claimed to have established with new statistical analyses is certainly an underestimate. These are grave accusations that the AD herring test was deliberately used for economic gain.

First, we describe the novel aspects. The authors went back to the original data, which includes verbal descriptions of a sentence or two written down just after the visit, at which point a "provisional score" was given. They convert this verbal report into a score (on a six-point scale) for the herring taste, subjectively, but using verbal guidelines which they formulated themselves. (They gave the guidelines to four independent persons who "replicated" the grading process; it seems fairly stable.) Now they used the taste variables *maturity* and *cleaning*, plus their new construct, to predict the *provisional score* by a new linear model. They discovered firstly that there is still an appreciable error term, and secondly that the Atlantic outlets got just significantly higher scores, when the *Atlantic* variable is included in their model.

Their damning conclusions are based on their new tripartite assumption: the provisional score should *only* be based on taste; *all* relevant aspects of taste have been expressed in the available scores and the verbal report; and their final model *perfectly* accounts for all of them.

Yet, the jury was out to *rank* participating herring outlets for the benefit of the public. The consumer does not want to eat in dirty establishments with unfriendly servers. The consumer does not want to get sick from eating the product. The experience of going out to eat Dutch new herring depends not just on the *set* but also on the *setting*. All three fundamental assumptions of Vollaard and Ours[5] are manifestly wrong. The AD herring test was never "only about the taste".

Aftermath

The publicity generated by Vollaard's two 2017 working papers forced AD to discontinue its annual herring test. The senior testers and their families suffered intensely from the accusations that they felt were unwarranted. Senior management at both AD and Atlantic felt their reputations had been damaged, as well as that of their businesses. AD fought back with an accusation that Vollaard's work violated standards of scientific integrity.

Tilburg University's department for handling such complaints – and, on appeal, that of the relevant national agency – concluded there had been no scientific integrity violation, but did state that further scientific debate was needed. Vollaard renewed his claims.[5] We analysed old and new arguments.[4] We notice the same failure of basic statistical assumptions for drawing conclusions even just of a descriptive nature from linear regression models, and we notice that the dependence of scores on location is much more complex than the simple binary variable "more than 30 km from Rotterdam" allows. In fact, when we model spatial dependence in a more realistic way we discover on the one hand multicollinearity due to confounding of the effects of space and of the *Atlantic* variable, and on the other hand we see evidence of an intuitively plausible spatial dependence (see Figure 2) on distance from the west coast of the Netherlands. People like to eat fish when on their seaside holiday, and this coast is where huge numbers of Dutch and German tourists go for summer holidays and day trips. The beautiful ancient city of Maastricht in the deep south-east of the Netherlands, on the other hand, is justly famous for Burgundian gastronomical experiences; it is not the place one goes for a Dutch new herring.

It never became apparent to Vollaard that correlation does not indicate causal relations unless very stringent assumptions are validated, nor did he notice that "borderline statistical significance" can be an artefact of model misspecification.

After the termination of the AD herring test, a newspaper based in the nearby city of Leiden started its own herring test. The panel of tasters was now a collection of 15 celebrities, who tasted herring blind (and therefore, not on location). It was initially a local affair but over the years it was

The senior testers and their families suffered intensely from the accusations that they felt were unwarranted

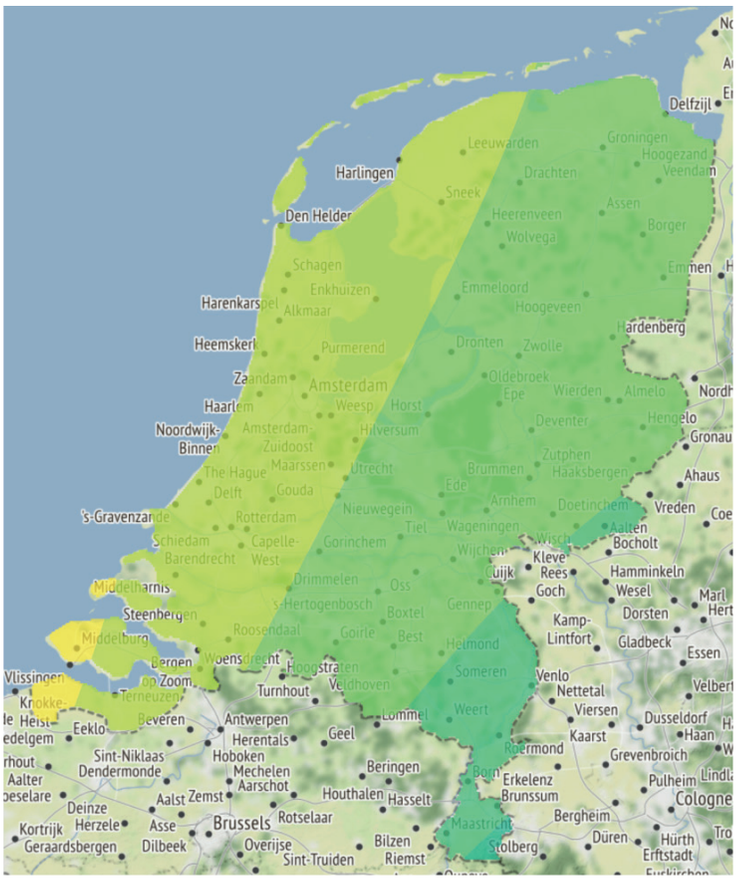

Figure 2: Spatial effect in fishmongers' scores modelled by quadratic spatial terms. Here, we run the linear regression almost identical to that in Vollaard's first paper,[2] without the *k30* variable and with the addition of the following variables: the longitude and its square, the latitude and its square, and the product of the longitude and latitude of each vendor's geographic location. The differences in score associated with the spatial terms in the regression model obtained are plotted, with the scores shifted such that the centre of the Netherlands corresponds to the score 6. The *p*-value for testing whether these quadratic spatial terms have an effect in the model is of the same order as that of *k30* in Vollaard's paper.[2] Recalling the clustering of the Atlantic outlets on the coast in Figure 1, we remark that the reported spatial effect offers an innocent (confounding) explanation for both of Vollaard's accusations.

expanded to become a national "herring taste test" with 25 new celebrities forming each year's panel. However, after a few years the newspaper found it increasingly difficult to recruit new herring outlets for the test, and it never amassed the level of national popularity and keen interest achieved by the old AD herring test. It was abandoned, though soon "rebooted" by a commercial organisation running big, public, mediagenic events. It remains to be seen how successful this will be.

A reckless rush for impact?

With the advent of modern software, it is easy and tempting to fit pro forma linear regression models to quickly generated data sets. Though it can be a good way to start exploring a data set, reading much into the regression results without careful examination of the models and data is like trespassing in minefields. Statistics are only as good as the model which produces them, and often such a "first try" model is simply not good enough. It is foolish or even reckless to insinuate causality based on statistical significances from a simple linear regression model. In this example, fishmongers are not a random sample from a well-defined population, but are a self-selecting group, from which drawing causal conclusions is audacious. Even drawing sensible descriptive conclusions is not trivial. Vollaard and his university PR department yielded to the temptation for rapid societal impact, causing a huge amount of damage.

Statistics are only as good as the model which produces them, and often such a "first try" model is simply not good enough

Statistical scientists must hold themselves to a high standard. Of course, there is a place for self-publication of exploratory statistical analyses of data concerning societally important questions. However, such analyses can have an immediate economic and/or political impact, so peer review is essential – not the formal peer review of a journal publication, but the usual process of scientific debate through seminars, lectures, and, most importantly, sharing of data and models at an early stage.

Anything less leaves a bad taste in the mouth… ■

Declaration of interest

Richard Gill initially analysed Ben Vollaard's statistical methodology as a consultant to the newspaper *Algemeen Dagblad* during a 2017 scientific integrity investigation carried out by the Dutch National Agency for Scientific Integrity (LOWI).

References

1. The Economist (2017) Netherlands fishmongers accuse herring-tasters of erring. *The Economist*, 25 November. bit.ly/42Tg5Ls

2. Vollaard, B. (2017) Gaat de AD Haringtest om meer dan de haring? Tilburg University. bit.ly/3qTxYMM

3. Vollaard, B. (2017) Gaat de AD Haringtest om meer dan de haring? Een update. Tilburg University. bit.ly/3XmRduf

4. Gao, F. and Gill, R. D. (2023) Pitfalls of amateur regression: The Dutch New Herring controversies. *Scandinavian Journal of Statistics*. https://doi.org/10.1111/sjos.12662

5. Vollaard, B. and van Ours, J. C. (2022). Bias in expert product reviews. *Journal of Economic Behavior & Organization* 202: 105–18.