|  |
| --- |
| **Using Data Badly – a user’s guide for the unwary geography student** |

**1. Introduction**

*How to Lie with Statistics* is a best seller within the statistical literature. Authored by the writer Darrell Huff, it is a short text, first published in 1954 and still in print today. Within it, Huff shows how statistics routinely are misused, misunderstood and misinterpreted in businesses, in public policy, in journalism and in other ways. In writing the book, Huff’s intention was not to encourage falsehood. Quite the opposite: he was teaching people how to avoid errors by exposing the dodgy applications. More recent books explore a similar purpose. For example, *The Tiger That Isn’t* by Andrew Dilnot and Michael Blastland (2007) and *A Field Guide to Lies and Statistics* by Daniel Levitin (2016). Understanding the validity of statistics is important because they shape our beliefs and reinforce our prejudices, especially in an era when they can be circulated at speed on social media with little regard to their veracity. Data skills allow us to do what Huff described as talking back to statistics – to have the knowhow that does not accept every statistic at face value but can spot errors, misunderstandings and brazen manipulations of the truth.

What has this to do with geography? The answer is that data skills are essential in geography. They are central to how we practise geography, to how we acquire geographical knowledge, and to how we apply geographical concepts and ideas to particular studies and contexts to facilitate geographical learning. They are skills that will aid the successful completion of a geographical project and independent investigation. However, they are also hazards that can ‘trip up’ the unsuspecting student. Fortunately, the most common types of problem can be avoided with a little forethought and care. The aim of this article is to provide guidance.

**2. An overview of the hazards**

In broad terms, the process of using data in a project can be looked at as the four steps shown in Figure 1. These are:

1. to think of question – a research hypothesis – that is in some way answerable using the data available or which could be collected;
2. to collect or to obtain the data;
3. to analyse the data to obtain an answer; and
4. to present and to communicate that finding to others in ways that are clear, honest and understandable.

These steps will be familiar to geography teachers and will form the basis of fieldwork and their students’ A Level individual investigations.

In practice, the actual process of completing a project may be less orderly, moving backwards and forwards between the steps. For example, it is not uncommon for new ideas to emerge from the data, changing the focus of the study and the research hypothesis. Even so, at each step there is potential for things to go wrong either by error or by design – by asking the wrong question, by collecting poor data, by misanalysing the data or by presenting the results in a manner that is either inept or plainly dishonest. The rest of this guide includes some cautionary tales.

Figure 1. Four steps for using data in a research project. If any one of these goes wrong the work will suffer.

**3. Asking the wrong question**

*“Where do you see yourself in five years?”*

*“In a mirror”*

Okay, it’s a weak joke (and not even original – I found it online) but it makes a useful point: if you don’t ask the right question (or if the question has an ambiguous meaning) then you may not get the type of answer you are looking for. Two examples are illustrative. Sadly, the first is tragic.

The NASA Space Shuttle, Challenger, disintegrated shortly after lift-off in 1986, killing its crew of seven. The disaster occurred because of a seal failure on the booster rockets. Concerns about launching were discussed the evening before because of known design flaws and because of the unusually low weather temperature expected on launch. However, data from previous launches showed that when damage occurred to the seals, it did so across a range of temperatures, suggesting temperature, of itself, was not a cause.

Unfortunately, as the statistician David Hand notes in his book *Statistics: A Very Short Introduction* (2008), this is a case of drawing the wrong conclusion from posing the wrong question. Aside from the fact that the shuttle had not launched on such a chilly day so existing information was not necessarily reliable, the data did not include the number of times the shuttle had launched without damage to the seals. Why did that matter? Because the important question is not whether the shuttle can (in principle) launch across a range of temperatures but whether the risk changes – are there more seal failures to non-failures at colder temperatures than at warmer ones? In fact, the damage free launches tended to be on warmer days so the much colder temperature increased the risk; fatally so.

The second example involved no loss of life but did involve a loss of money and of reputation (for a while, at least). It is notorious in business schools and concerns the launch of ‘New Coke’, replacing ‘Classic Coke’ in the mid-1980s. A motivation for the revised product was Pepsi’s increasing sales eating into Coke’s market share, which in turn was driven by a successful marketing campaign, including the Pepsi Challenge. This was a blind tasting of the two drinks (so the taster didn’t know which was which) given to people in locations such as shopping centres, which appeared to show an above average preference for Pepsi. However – and as the author and journalist Malcolm Gladwell observes in his book *Blink* (2005) – the ‘sip test’ is not actually an accurate test of consumers’ buying preferences because what is liked on an initial taste (Pepsi) is not necessarily what is preferred over the course of a bottle or can. Furthermore, although Coke could run lab tests establishing a preference for New Coke, the response “I prefer the taste of product X” in an artificial situation is not the same as “the next time I am in a supermarket, I will buy it instead of the traditional product, which I am happy for you to stop selling.” Buying behaviour is not motivated solely by (blind) tasting but by other factors such as branding and attachment to a product. In the late 1990s, the vice President of Marketing who had led the initiative, Sergio Zyman, summed up the New Coke experience: “Yes, it infuriated the public, cost us a ton of money and lasted for only 77 days before we reintroduced Coca-Cola Classic. Still, New Coke was a success because it revitalized the brand and reattached the public to Coke.“[[1]](#footnote-1)

Of course, a student project is unlikely to suffer multi-million-pound consequences. Even so, giving care to the research question is important. A frequent problem is for it to be too wide, too vague and sometimes too ambitious for it to be answerable within the constraints of the data that could realistically be obtained given the time and resources available. A question such as “what is the future for nuclear power as a source of domestic energy supply in the UK and how does it compare to renewable sources?” may well be interesting but it is very hard to answer in any definite way. The question is too wide, in fact it consists of two questions (“what is the future…?”, “how does it compare…?”) and, because it is speculative, it is likely to result in vague conclusions. The question, “do older people in my community have a different attitude to nuclear energy than younger ones?” is far more manageable and can result in an answer (measure the attitudes in some way, group the measurements by the respondents’ age categories and look for differences).

Whether it’s a reliable answer depends upon the quality of the data and how they are collected, which are important issues to be returned to presently. For now, the immediate point is to make sure that the research question is framed in a way that allows for an answer. Even if it is, there is still a potential gap between the researcher and the researched in terms of the way a specific question is asked and understood, which is what the mirror gag is illustrating. This is an issue especially when using questionnaires and surveys, which is why undertaking a small pilot study to see how the questionnaire ‘works’ and to allow for any changes is a good idea before sending it out more widely.

**Checklist for doing an individual study**

Has your research got a question that is not-too-vague or wide-ranging?

Can the question be answered in some way within the constraints of time and of the resources available to you?

**4. Obtaining unreliable data (or too little, or ‘too much’)**

Electoral forecasting, which is to predict the results of an election from opinion polls of the voters, has come under scrutiny recently. The polls failed to predict the Conservative victory in the 2015 election, the Brexit vote in favour of the UK leaving the EU, and Trump’s victory over Clinton. To be fair, the previous statement is not entirely true: arguably they did, in the sense that all predictions come with a margin of error (a range of possible outcomes) and those margins typically included what transpired. Nevertheless, there was something peculiar in how most of the polls seemed repeatedly to be pointing towards the incorrect results.

It’s instructive to reflect on what the polling companies are trying to do, which is to predict the voting intentions of the whole electorate from a survey of a few of them, typically about 1000 to 2000 people. Statisticians call those people the sample; the electorate is the population. The population is not necessarily everyone who lives in a country and, in other circumstances, need not be people at all. The population for a sample of river contaminants is the river. The population is that which we are trying to obtain information about. In the present example, it is everyone who could vote. The sample provides the data but it is the population that is of real interest.

How can this work? How can a sample of a few thousand people provide reliable information about the voting intentions of about 46 million people (in the UK, in 2016)? [[2]](#footnote-2) A common mistake is to worry about the seemingly small sample size. That’s not the issue; it is sufficient. About the same number is used in the US to survey an electorate of hundreds of millions. What matters is how the survey data are collected. There is a readable explanation about how it is done at <http://ukpollingreport.co.uk/faq-sampling>. What it boils down to is trying to ensure the sample is representative of the population.

In statistical terms, the aim is to ensure the sample is not biased. Bias means the sample is systematically misunderstanding the population in some way – for example, by surveying only younger voters or by taking water quality samples only upstream of a sewage discharge. A common way to avoid bias is to take a random sample of the population of interest. That doesn’t prevent any one sample from being misleading. Precisely because it is random it could (by poor fortune) select only young voters to ask. However, it does mean that if the sampling could be repeated over-and-over again, averaging over the samples would yield measurements representative of the population.[[3]](#footnote-3)

In practice, polling companies focus on particular places or on panels of people who are willing to be interviewed to obtain their predictions, deliberately ‘over-sampling’ and subsequently making a statistical correction for small population groups. (If, for example, you sampled 1000 voters truly at random it’s quite likely all of them would be White British, which is a concern if voting intentions vary by ethnic group). In the context of a student project, collecting truly random data might also be very difficult. However, that doesn’t prevent an effort for at least a quasi-random approach; for example, measuring (e.g. vegetation) at regular intervals along a transect or interviewing every 10th person who walks past. What you should be cautious about is a convenience sample – one where the sample is determined solely by ease of access. These include Facebook or other social media polls of friends and family. Sometimes it makes sense to use a convenience sample if you have access to a group of people and it is those people that your study is interested in. The major drawback is if you attempt to generalise beyond those people and assume the sample is representative of a wider population. Overly convenient samples can also give the impression that you have not tried very hard to collect the relevant data.

At a minimum, it is important to be aware of the potential biases and limitations of data when attempting to draw conclusions from them. Don’t do what the *Daily Express* has done on at least two occasions, which is to generate nonsense statistics to create headlines:

*99% of you say: Get us out of Europe* (November 26, 2010)[[4]](#footnote-4)

*No delay on EU exit: PM reacts as 99% of Express readers say they feel 'betrayed'* (November 5, 2016)[[5]](#footnote-5)

Any survey that generates 99 per cent support is immediately suspect and, in this case, those who have the motivation (and the time) to ring-up and lodge their opinion are likely to be a biased sample even of Daily Express readers. Allowing that the survey is not random but self-selecting, the 99 per cent figure remains implausibly high. It may be that what the newspaper asked is for those who support its campaign to leave the EU to ring up and register their support. In that case, the correct headline would be *99% of you who rang to support our campaign to get us out of Europe say, ‘get us out of Europe’*. If so, it makes you wonder what was going on with the remaining 1 per cent (perhaps they were calling for a pizza?). The data are highly biased.

Assume yours won’t be; that you have good reason to suppose that the data will be representative and fit for purpose. How much data do you need to collect? The answer depends upon how variable the population is. If you could anticipate in advance that everyone would vote the same way then you would need to ask only one person. If voting intentions vary by age, gender, social class, location, ethno-cultural background, level of education, and so forth then you need to capture this diversity. It also depends upon how secure you want to be in your findings.

As a rule of thumb, a sample size of about 30 is often considered the minimum if you want to undertake a traditional statistical test with the data. Below this, the statistical test lacks what is known as power: it is hard to obtain a statistically significant result because the sample size is too small to find anything within it. It’s not a hard-and-fast rule though. It would be unwise to allow the sample to be any smaller but it will need to be larger if your analysis splits (cross-tabulates) the data in some way. Consider, for example, a survey of 30 people that looks for differences in an attitude by (conventional categories of) gender. All things being equal, there will be about 15 people in each gender group, which is not so many after all. Sub-divide that by ethnicity, for instance, and you may find there are none or one person in each group, which is clearly insufficient if you are interested in the differences between those groups.

It may be noted that qualitative research sometimes draws on smaller numbers when undertaking focus groups, for example (which, in a political context, could be to see how attendees respond to differing styles of political advertising; or, in a University context, to lean from the recipients about how the provision of a bursary has supported them through their studies[[6]](#footnote-6)). However, in these circumstances the attendees will be purposefully selected to represent either mixed or specific points of views. The purpose of a focus group is also different as it is a less structured (a more open) exploration of people’s perceptions, opinions, beliefs, ideas, values and attitudes than is a typical quantitative survey. Qualitative research offers the space for responses to be made that would otherwise be over-looked or are difficult to measure – indeed, they can challenge the researcher’s prior understanding of the subject and reframe the nature of the study. Nevertheless, it would be naïve to believe that issues of representativeness no longer apply if, at some point, the study intends to claim that what is discovered from the focus group applies beyond those who are present to reflect wider experiences and sentiments of those who are not.[[7]](#footnote-7)

A further consideration is whether you need to collect the data at all. Data collected by a person herself are known as primary data. Secondary data are those collected, often by government agencies, for a purpose other than the research project but that may nevertheless be useful for it. Indeed, those data may offer far greater scope for analysis than anything that could be collected by the researcher alone. Investigating how political attitudes (e.g. support for coalition governments) vary between voters in England, Scotland and Wales would be difficult if you had to collect the data yourself. It is easier to use the British Election Study’s Data Playground: <http://www.britishelectionstudy.com/data-playground>.

There are lots of secondary data sets, including the [UK Census](http://www.census.ac.uk), [survey data sets](https://www.ukdataservice.ac.uk), [climate data](https://www.metoffice.gov.uk/public/weather/climate/), [hydrological data](https://nrfa.ceh.ac.uk/), [world development data](https://data.worldbank.org/data-catalog/world-development-indicators) and very many others – search online but make sure the data are supplied by a reputable source and be aware of any possible limitations. Consider the geographical scale of the data set, when it was collected, why it was collected, for whom and why.

You can have ‘too much’ data. With a large data set (of 10,000s, 100,000s or even millions of records) almost every statistical test produces a result that is ‘statistically significant’ even if what is being tested for (the average difference between two samples or the relationship between two variables, for example) is trivially small. The problem is that statistical significance is related to the sample size so the more data you have the more likely it is that the test will return a statistically significant result.

In fact, the statistical tests you are most likely to encounter at school were not really designed to cater for large data sets. It would be daft to throw away good data just to make the sample smaller (you can’t learn more from less) and silly to collect your own data when there are lots of secondary data sets available that are ideal for geographical work. Instead, keep in mind the difference between *statistical* significance, which is a statistical concept, and *substantive* significance, which is about whether what you have measured is of any real-world consequence or not. To illustrate this, consider that the relationship between two variables can be measured as a correlation that ranges from –1 (a perfect negative correlation; e.g. between a variable, *x*, and another equal to 1–x), to +1 (a perfect positive correlation; e.g. *x* and 2*x*) through 0 (no apparent relationship). A correlation of 0.01 is therefore somewhat trivial (it is very close to zero) but given enough data it can still be statistically significant.

Some people have been critical of the whole idea of statistical significance testing – the scepticism is apparent in the title of Stephen Ziliak’s and Deirdre McCloskey’s (2008) book *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives* (the standard error is an ‘under the hood’ statistical measure used in significance testing). The point is that it’s not enough just to take some data and apply a statistical test to it in a robotic manner. You need to interpret the result – to step back and ask what it really means and whether it matters (and, if so, for whom, and why?). It’s that act of careful thinking that leads from data to useful knowledge.

**Checklist for doing an individual study**

If you are collecting data, how can you do so in a way that avoids overt bias and helps ensure that the information you obtain is representative of that which you are studying?

How much data do you need to collect? Is there potential for using existing (secondary) data instead?

**5. Mis-analysing the data**

In 1998, a paper – subsequently retracted – was published in the medical journal, *The Lancet*. It purported to show a link between the combined measles, mumps and rubella (MMR) vaccine given to children and autism. It caused a media sensation and a public health scare. Subsequent investigation has described the paper as fraudulent but it has left a legacy of suspicion against the vaccine and consequently a decreased uptake, creating a situation that is more dangerous to health than any risk from the vaccination (the scientific consensus is that the MMR vaccine has no link to autism).

More recently, in 2015, the carmaker, Volkswagen, was found to have manipulated car emission tests so that its vehicles were programmed to behave differently under testing than for real-world driving. Other car companies are being investigated too. At the time of writing (September 2017), the foreign secretary, Boris Johnson, is embroiled in a spat about the integrity of statistics and their use. He has been reprimanded by the UK statistics watchdog for resuscitating (in a newspaper article) the controversial claim, much promoted by the Leave campaign during the Brexit referendum, that Britain pays £350 million a week to the EU. That exaggerates the true value of about £250m by 40 per cent. When money paid back from the EU is taken into consideration, it is nearer £120 to £160 million.[[8]](#footnote-8)

You would not (I hope) resort to falsifying data nor using statistics selectively. Even so, the potential for errors to enter and to undermine your work remain. Those are unlikely to be exposed by the glare of the media’s spotlight but an examiner might notice and mark your work accordingly. We have discussed already the problems of biased and unrepresentative data, of collecting too little and of interpreting statistical tests without consideration to their real world meaning. Let’s consider some other examples of where things can go wrong.

One problem can arise from the type of data collected. In broad terms, data can be continuous, meaning they can take on any value within a given range (and to a certain number of decimal places, as determined by the precision of the measuring instrument). Daily temperature, for example or river water levels. Alternatively, they can be discrete, which means the range of possible values is limited to a small number of usually whole values (with no decimal place). Traffic accidents at a road junction, for example (typically zero, sometimes one, occasionally two, infrequently three, at times four, … but never 1.5 or any other fraction of an accident). They can also be categorical (sometimes described as qualitative) where the numbers may be arbitrary (1 = males; 2 = females) or have a rank order (1 = low level of child poverty; 2 = middle level of child poverty; 3 = high level).

The analysis that can be done with the data is dependent on the data type. You can calculate the mean (average) of continuous data but for categorical data it is nonsensical (what would it mean to say that the average respondent to a survey has a gender of 1.4?). It is, however, possible to calculate the frequency of each category in the data and to determine which is most prevalent. If there are 100 males (60%) and 67 females (40%) then males are the modal (most frequently occurring) category. Doing the same for continuous data is foolish because almost every value will be unique. But continuous data can be organised into groups (for example, temperature bands: 0 – 9.9°C; 10 – 19.9°C, etc.) and the frequency of each group determined. When those frequencies are plotted visually they produce a histogram; the equivalent chart for categorical data is a bar chart: <http://stattrek.com/statistics/charts/histogram.aspx>.

Imagine you are interested in the relationship between two variables – neighbourhood unemployment rates and neighbourhood crime rates. The (Pearson) correlation coefficient is suitable for these two continuous variables but if the neighbourhoods have been ranked by their levels of unemployment and, for the second variable, by their levels of crime then the Spearman’s rank correlation should be used instead.

You can always convert continuous data into ranked or categorical data but you cannot go the other way around. Continuous data are usually better because they contain more information and more can be done with them. However, there may be pragmatic reasons to prefer other data types. In surveys, for example, it is often easier for the respondent to have a list of categories to choose from than to try and provide an exact value of income, for instance, or the number of journeys travelled by public transport over the last week.

If you are using correlation to explore the relationship between two variables then be wary of correlating the counts for one variable with the counts for another, especially when using census data about neighbourhoods. If you do, you may well find, for example, that the number of people unemployed per neighbourhood increases (is positively correlated) with the number of people not owning a private vehicle. Since this fits with intuition – it takes a certain amount of money to buy and to maintain a car – you probably wouldn’t question the result. You should. Census neighbourhoods vary in population size. Regardless of any actual relationship between unemployment and car ownership, any areas containing a lot more people are likely to have more people unemployed and more without a vehicle due to the population size alone. If there are a lot of people then there are more to be unemployed or not own a vehicle. Potentially, all the correlation is really telling is that higher population areas contain more people than smaller ones.

The way to resolve this problem is to better compare like-with-like, which is achieved by correlating the unemployment rate (e.g. percentage of the neighbourhood population that is unemployed) with the corresponding vehicle non-ownership rate. It is important not to confuse counts with rates. It is entirely possible for the number of people in employment to increase but for the unemployment rate to increase also. That would happen if the number gaining jobs increases but the number of people looking for jobs also increases and more greatly. It is not uncommon for politicians to pick and choose between counts and rates in accordance to which conveys the most positive message.

Assume that this error has been avoided and that there is indeed a positive correlation between the unemployment and vehicle non-ownership rates. Owning a car requires money so it make sense. But does unemployment lead to vehicle non-ownership or is it the other way around, since reaching a workplace can be harder without a car? Of itself, the correlation value provides no answer. A correlation coefficient is a measure of whether one variable can be used as a substitute for another, which is the case if the two variables have some sort of relationship (as one increases, so too does the other, a positive relationship; or, as one increases, the other decreases, a negative relationship). What it does not prove is causation – that the increase/decrease in one variable is the cause of the increase/decrease in the other.

In discussions of social mobility in the UK, it is sometimes claimed that it stalled after many of the former, academically selective grammar schools were converted into comprehensives. Perhaps so. However, coincidence of timing does not mean that they are directly related. Much has changed economically and socially since the 1960s and 1970s, any of which could be the true cause. Equally, critics of foreign aid spending may note that as it increased in Africa as a percentage of national GDP especially during the late 1970s to mid-80s, growth per capita in African countries slumped.[[9]](#footnote-9) Does that mean that the aid caused the slump? Possibly. But that was also a period of oil price shocks, economic restructuring and the ascendancy of neo-liberal economic policies worldwide. Those or other determinants could be the true causes. The website, <http://www.tylervigen.com/spurious-correlations>, contains some quirky examples of spurious correlations. For example, per capita cheese consumption correlates with the number of people who died by becoming tangled in their bedsheets.

The take-home message is not to jump to conclusions nor extend them beyond what actually has been revealed. That warning applies to another sort of error that arises from subtly moving the goal posts in terms of what is being analysed. If there is a correlation between the unemployment rate and the vehicle non-ownership rate for neighbourhoods does it mean that the unemployed are less likely than the employed to own cars? You might think so but it’s not actually proven: unemployed persons could be living in places where other people choose not to own cars. The error is to assume that a relationship measured at one scale, neighbourhoods, must necessarily apply to another scale, individual people. Formally, it is known as the ecological fallacy. It is a potential problem for commercial classifications of postcodes into neighbourhood types that are then used to direct advertising or to plan for businesses and services. One product has a group of neighbourhoods described as ‘Low Income Workers - Older social renters settled in low value homes in communities where employment is harder to find.’ Does that description apply to everyone in those neighbourhoods?[[10]](#footnote-10) Probably not.

A final consideration is that statistics such as Pearson correlation values and others such as mean averages are susceptible to outliers – that is, measurements that are different by having very high or low values in relation to the rest. A somewhat hackneyed example is to imagine measuring the annual income of a group of random people that happened to have Bill Gates in the room. The average for this group will be deceptively high.

Can anything be done to prevent such anomalies? Giving consideration to how, where and when the data are collected will help but may still result in unusual measurements. Consequently, taking time to check the data is important through numeric and graphic summaries that reveal the centre (the average) and the spread (e.g. the range) of the data. The Bill Gates example illustrates why averages should be used with measures of variation or spread around the average: the range, the interquartile range, variance or standard deviation, for example. Histograms and box plots (also known as box-and-whisker plots) can help to detect unusual values.

Some statistical measures have counterparts that are robust to outliers. The median (the middle value) is a better way of calculating the typical income in a country than the mean is. The middle is the middle regardless of how extreme some very high incomes may be. However, by raising the mean, the presence of those high earners gives a misleading impression of what the average person earns.

Omitting any anomalies from the statistical calculations is a sensible precaution to see how much calculations change as a result. If they don’t change greatly then there is not much of a problem. If they change enough to affect your interpretation of the data and of the results then that is more of a concern because it implies that what you were observing was driven by one or a few unusual cases. However, just because they are unusual doesn’t make them wrong. They could be an error but they could also represent interesting cases that merit further study. The geographical questions are what leads them to be different and is there a geographical explanation for it?

**Checklist for doing an individual study**

What statistical techniques do you anticipate using?

Does the type of data you will be using support the statistical analyses you want to do with the data?

What factors will you have to pay attention to when drawing conclusions from the data?

**6. Miscommunicating and mispresenting the results**

Data skills do not begin with the collection of data nor end with the analysis of them. They include communication of the results and conclusions in ways that are honest and do not distort the meaning of the data. Graphs and figures play a key role in this.

Graphics have a visual appeal that can be very persuasive. Unfortunately, that means they can also misinform. I’ve featured elsewhere my ‘favourite’ example; it is in the Short Introduction to Quantitative Geography on the RGS-IBG website. It bears repeating and is shown below, having originally appeared on the website of a national UK website. There is an obvious bias that is clear from the title. What is less obvious is that the “spectacular miscalculation” claim is untrue and, what is more, the graph shows it. By definition, a 95 per cent estimate will not be correct 100 per cent of time so even if the world average temperatures did “crash out” of the predicted range that would not immediately invalidate the predictions. However, the temperatures don’t fall outside the predictions so a more accurate headline would be ‘*graph reveals how 95 per cent certain estimates of the earth heating up are correct.*’ Admittedly the graph was circulated in 2013 when there was some suggestion of a pause in global warming but by 2015 the change in temperature had risen by 1 degree, as predicted by the forecasts but now at their upper end.

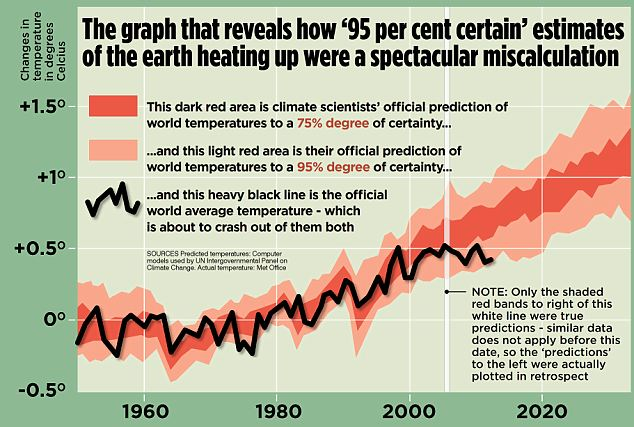


Figure 1. An example of a misleading graph (source: <http://www.dailymail.co.uk/news/article-2294560>)

Most misuses of graphics are subtler than this. Pie charts are often used but they are generally regarded by people with interest in visual communication as a poor way of presenting data, the main problem being that pie charts work on angles but angles are hard to interpret and to compare with one another. Bar charts are often better for presenting results. However, bar charts can be misused too. A popular trick is to exaggerate differences by choosing a non-zero base on the y-axis. Figure 2 shows passenger numbers at five major airports in 2012 and 13. Looking at it you may think that the numbers in Atlanta (ATL) are about five times that of Chicago (ORD) but that’s not true. They aren’t even double. The visual deception is created by having the base of the graph at 60 million instead of zero. Changing the base emphasises the differences; however, it also exaggerates them. For another example showing how political parties use this technique to encourage votes, see the video by the Sheffield Methods Institute at <https://www.youtube.com/watch?v=hYaoE4Kh9fk>.

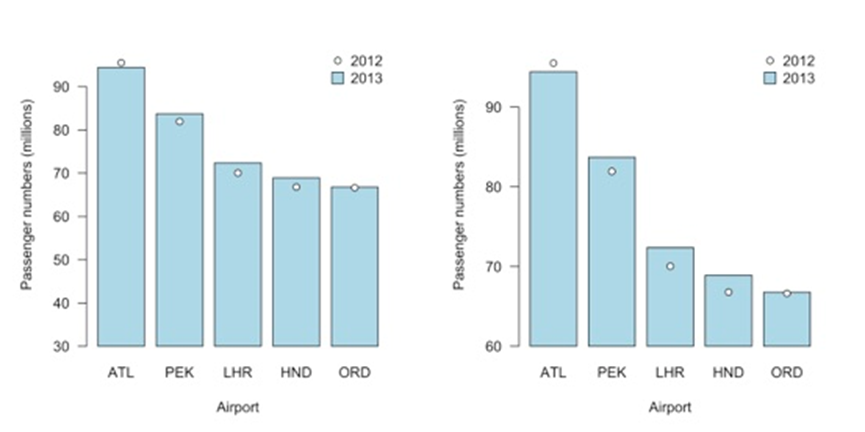


Figure 2. Another example of a misleading graph (source: Harris, 2017)[[11]](#footnote-11)

Changing how the data are presented will affect the interpretation of what those data are saying. There is a geographical example also by the Sheffield Methods Institute that shows how housing prices in London can be used to illustrate two different points of view dependent upon how the data are mapped. [This video](https://youtu.be/G0_MBrJnRq0) is worth watching although it also contains a slight (but common) error. Look very carefully at the map displayed [here](https://youtu.be/G0_MBrJnRq0?t=46). The map classes (the groups of values used to shade the map) go from £0 to £1871711, then from £1871711 to £3743421, £3743421 to £5615132, and so forth. What’s the problem? Which group is a value of £1871711 in – the one from £0 to £1871711, the one from £1871711 to £3743421 or both? It’s a relatively minor detail but it could be avoided: £0 to less than £1871711; £1871711 to less than £3743421, etc.

Each week, this [website](http://www.makeovermonday.co.uk/) posts a link to a chart, and its data, and invites its followers to rework the chart, either to improve it or to present new stories from the same data. [This example](https://trimydata.com/2017/09/24/mm-week39/), is a makeover of a chart showing how dietary requirements vary around the world. It comes with a critique of the original chart and what has been done to improve it. You could compare the two and see if you agree with the changes.

An alternative to using graphs is to present the data in a table. These can be effective, especially when the numbers need to be communicated more precisely than a graphic allows.   
The disadvantage is that tables are not as visually appealing, can take longer to interpret and can become very long and unwieldy if there is a lot of information to present. They can also look slapdash and rushed. Table 1 is unattractive with its black header column, thick borders and unnecessary stripes. It also lacks consistency: the percentage changes are shown to a varying number of decimal places. Little consideration has been given to the row order and what the reader should learn from the data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Airport | City | Passengers 2013 | Passengers 2012 | % change |
| ATL | Atlanta | 94431224 | 95513828 | -1.333 |
| HND | Tokyo | 68906509 | 66795178 | 3.16 |
| LHR | London | 72368061 | 70038804 | 3.32566 |
| ORD | Chicago | 66777161 | 66629600 | 0.22 |
| PEK | Beijing | 83712355 | 81929359 | 2.176 |

Table 1. An ugly table of the airport passenger data.

Tables 2 and 3 are much better. These two are the same except in the way the data are sorted. The first emphasises the busiest airport (ATL). The second emphasises the one with greatest yearly change (LHR). Neither is wrong or right but the choice does affect the take-home message of the data. The data are presented tidily and consistently, in all but one case using three significant figures but the same number of decimal places for each.

The decision to use three significant figures is a matter of personal taste. I could write that the percentage change in passenger numbers was -1.13345263473 per cent in Atlanta over the period but the precision is spurious and does not add to the key point, which is that passenger numbers fell slightly over the period. Arguably, rounding degrades the data. For example, the number of passengers for ATL is now recorded at 94.4 million (i.e. 94400000) when the true value is 94431224. Again, what matters is what we are trying to communicate. Rounding makes it no less obvious that Atlanta is the busiest airport. Furthermore, it’s much easier to appreciate the scale of the numbers when they are clearly reported in millions.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Passengers (millions) | |  |
| Airport | City | 2013 | 2012 | % Change |
| ATL | Atlanta | 94.4 | 95.5 | -1.13 |
| PEK | Beijing | 83.7 | 81.9 | 2.18 |
| LHR | London | 72.4 | 70.0 | 3.33 |
| HND | Tokyo | 68.9 | 66.8 | 3.16 |
| ORD | Chicago | 66.8 | 66.6 | 0.22 |
|  | Total | 386.0 | 381.0 | 1.39 |

Table 2. A better tabulation of the data, with the rows ordered to emphasise the busiest airport

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Passengers (millions) | |  |
| Airport | City | 2013 | 2012 | % Change |
| LHR | London | 72.4 | 70.0 | 3.33 |
| HND | Tokyo | 68.9 | 66.8 | 3.16 |
| PEK | Beijing | 83.7 | 81.9 | 2.18 |
| ORD | Chicago | 66.8 | 66.6 | 0.22 |
| ATL | Atlanta | 94.4 | 95.5 | -1.13 |
|  | Total | 386.0 | 381.0 | 1.39 |

Table 3. In this example the airport with greatest growth is emphasised

Poor presentation can be amusing, unless you are being marked down for it. Still, for some light entertainment you can enjoy the following pie chart from Fox News (Figure 3). It makes very little sense at all. Further examples of how this news channel uses graphics to manipulate statistics (intentionally or otherwise) can be found [here](http://www.businessinsider.com/fox-news-charts-tricks-data-2012-11). Other visualizations that make no sense can be viewed [here](http://viz.wtf/) (where you can also upload any examples that you come across).

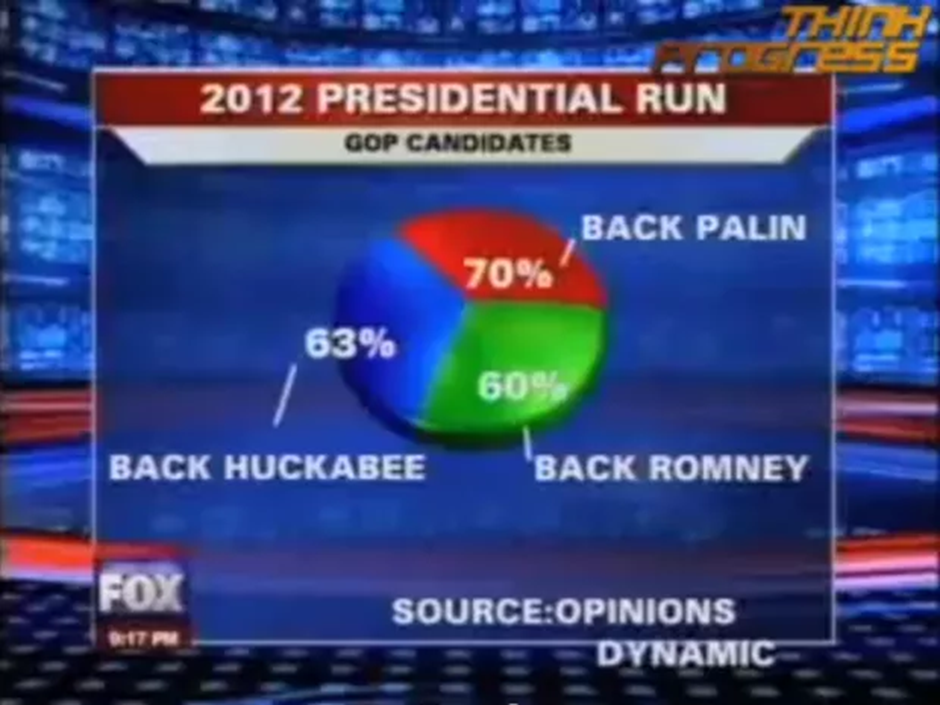


Figure 3. A very confusing graphic! (source: <http://www.businessinsider.com/fox-news-charts-tricks-data-2012-11>)

**Checklist for doing an individual study**

Presentation matters! Pay attention to how you can communicate your results in a way that is honest and does not distort the data. Simple graphs and well formatted tables are usually best. Avoid adding unnecessary visual clutter – for example creating a ‘3D’ pie or bar chart when the impression of depth merely makes the graph harder to read.

**7. Conclusion**

There is an excellent short guide to spotting a dodgy statistic on [The Guardian website](https://www.theguardian.com/science/2016/jul/17/politicians-dodgy-statistics-tricks-guide). It is worth reading as it serves as a useful *aide memoire* to the sorts of errors, mistakes and downright dishonesty that need to be avoided when using data skills and statistical techniques in geography.

If the potential to make mistakes seems a bit overwhelming then rest assured that what avoidance of doing so comes down to is working with honesty, integrity and with a bit of common sense, taking care to collect, analyse and present the data effectively to answer a clearly defined question. The answer may not be definitive because the same data can support multiple interpretations and points of view; for an example of this considering ethnic (de-?) segregation in London (paper on the RGS-IBG website). However, whilst we might debate the conclusions what we should not be debating is whether the research has been done well.

To support this and the application of data skills in geography, there is a range of online resources at the RGS-IBG Data Skills website: [www.rgs.org/dataskills](http://www.rgs.org/dataskills). Particularly recommended is a very helpful student guide to the A Level independent investigation. It is free to download from [www.rgs.org/nea](http://www.rgs.org/nea) and includes a guide to writing a research plan, a guide to effective background reading, a guide to forming a hypothesis (a research question), and substantial information on how to collect, present and analyse data.

**Acknowledgements**

This production of this guide was funded by the Nuffield Foundation as part of the RGS-IBG Data Skills project. The views expressed are those of the author and not necessarily those of the Foundation. A small part of it (under the section Miscommunicating and mispresenting the results) also appears in the book *Quantitative Geography: the basics*, which is published by Sage.

**About the Author**

Richard Harris is Professor of Quantitative Social Geography at the School of Geographical Sciences, University of Bristol where he is also the Director of the University of Bristol Q-Step Centre, part of a multimillion pound national initiative to improve the quantitative skills training of social scientists. In 2014, he was winner of The Royal Geographical Society (with the Institute of British Geographers) Taylor & Francis Award for excellence in the promotion and practice of teaching quantitative methods. He is author of several textbooks, including *Quantitative Geography: the basics* (Sage, 2016) and *Statistics for Geography and Environmental Science* (Prentice Hall, 2011; jointly written with Claire Jarvis).

1. <https://en.wikipedia.org/wiki/New_Coke> [↑](#footnote-ref-1)
2. <https://www.ons.gov.uk/peoplepopulationandcommunity/elections/electoralregistration/bulletins/electoralstatisticsforuk/2016> [↑](#footnote-ref-2)
3. That was what was odd about the polls about the 2015 Conservative election victory: they repeatedly seemed to indicate the wrong result, suggesting they were biased. [↑](#footnote-ref-3)
4. <http://www.express.co.uk/news/uk/213821/99-of-you-say-Get-us-out-of-Europe> [↑](#footnote-ref-4)
5. <http://www.express.co.uk/news/politics/728991/Theresa-May-pledges-delay-EU-exit-Brexit-High-Court-rulingst> [↑](#footnote-ref-5)
6. <http://www.bris.ac.uk/media-library/sites/geography/pfrc/pfrc1612-widening-participation.pdf> [↑](#footnote-ref-6)
7. Qualitative and quantitative approaches are sometimes presented as an ‘either – or’. However, some of the best research combines them both. [↑](#footnote-ref-7)
8. See <https://fullfact.org/economy/our-eu-membership-fee-55-million/>, <https://infacts.org/uk-doesnt-send-eu-350m-a-week-or-55m-a-day/>, <http://www.telegraph.co.uk/news/0/eu-referendum-claims-won-brexit-fact-checked/> and <https://www.theguardian.com/politics/reality-check/2016/may/23/does-the-eu-really-cost-the-uk-350m-a-week>. [↑](#footnote-ref-8)
9. Easterly W (2003) Can foreign aid buy growth? *Journal of Economic Perspectives*, 17 (3), 23–48. [↑](#footnote-ref-9)
10. <http://www.experian.co.uk/assets/marketing-services/brochures/mosaic_uk_brochure.pdf> [↑](#footnote-ref-10)
11. Harris R, 2017, *Quantitative Geography: the basics* (London: Sage) [↑](#footnote-ref-11)