Critique of Hsiang and Sekar (2016) NBER Working Paper 22314

Does legalization reduce black market activity? Evidence from a global ivory experiment and elephant poaching data

This paper aims to investigate whether the 2008 legal sale of ivory to China (and Japan although the focus is on the sale to China) is the reason for the increase in the illegal killing of elephants and the illegal ivory trade. The paper does this using data from two global monitoring systems for elephants – MIKE and ETIS – although focuses mainly on MIKE data – specifically PIKE.

I understand the key argument in the paper to be as follows.

You test whether there is a discontinuity in the PIKE data between 2007 (pre-sale) and 2008 (seen as post-sale). By discontinuity you mean that estimates of PIKE from 2003 to 2007 are significantly lower than estimates of PIKE from 2008 onwards, even after accounting for possible trends in the data. The paper describes a number of different models where most, but not all of the models, suggest that there is such a discontinuity.

You then look for a discontinuity between 2007 and 2008 in a number of other variables that you have selected to measure Chinese influence and presence in elephant range states. Because you don't find the same discontinuity in these variables the conclusion that you draw is that the discontinuity that you observe must be because of the sale – in other words the increase in illegal killing of elephants is due to the sale of ivory to China – because there is no other possible explanation.

My main criticism of this paper is the overall approach that you have taken and causal link that you make. However, even if I thought that the actual approach was valid I have criticisms of the statistical analysis of the PIKE data itself that describes the discontinuity. The trends in the PIKE data that you describe do not match those from other analyses which show an increase in trade from about 2006/7 onwards and do not show a discontinuity in 2008[1]. Finally, there are some major misunderstandings of the analytical approaches used to analyse the MIKE data in Burn et al (2011) and CITES reports and the ETIS data in Underwood et al (2013).

General approach

The logic presented in the paper is that an observed discontinuity in PIKE at the time of the sale can be attributed to the sale if other factors that might be related do not change discontinuously at the time of a particular event. This is claimed to hold true even if these other factors trend upwards or downwards, but do not change discontinuously.

This is an extremely weak strategy for measuring attribution. One might expect this strategy to work for a very simple system but the ivory trade is not at all simple. The trade is dynamic involving many different countries and players with different drivers operating at different places along the trade chain. Your strategy does not consider other complexities of the trade and the mass of other drivers that could be considered. For example, in the paper you do not:

[1] Although it is the case that these analyses have not primarily looked for these trends we have in the past considered a step change and not really seen any evidence for this.

- give an explanation regarding your choice of other potential drivers of the trade. I think there are a number of other things that could be considered that you have not looked at all.
- discuss the global financial crisis of 2008/09. Could this not be considered an event that could cause a discontinuity in demand at around the same point in time? How can the effects of this be disentangled from impacts of the sale?
- talk about trade in other illegal wildlife products that also show an increase in demand over the same time period - for example, rhino horn or pangolins where legal sales have not been permitted.
- consider trade in other goods that maybe have a similar role to ivory within China. How has demand in these changed over the same time period?

I would also question whether there really is a discontinuity in the data. It is possible that the data also support a model where PIKE is increasing smoothly over time rather than there be a step change. For example, you haven't shown the results of any fitted models which allow for different types of non-linear trends in PIKE – you have only allowed linear trends both before and after the sale. These models supporting different hypotheses could then be compared.

The paper assumes that the discontinuity is entirely attributable to the sale. It is a very strong assumption that there is only one driver of the trade. It is more likely, given the complexity of the trade, that if the sale has had an effect it will have contributed to a change in PIKE rather than be the only reason for an increase, for example see my point (c) above, and so it is a question of relative contribution rather than attribution that should be explored.

I would therefore want to look at the combined effect of many possible drivers rather than looking at each of them separately as you do here. It could be that some combination of drivers does explain any modelled discontinuity or trend in the trade. Such an analysis should try to directly link up trends in PIKE with potential drivers rather than looking separately for similar patterns in different sets of data. This is not at all easy and because of the potential for drivers working on different parts of the trade chain a complex modelling approach is required.

Statistical modelling

Even if your approach was OK the analyses that you have carried out are not.

The analyses of the PIKE data that you present in this paper does not follow the approach carried out in Burn et al (2011) and the trends do not look the same. I note the two main problems with your analysis.

PIKE data are proportions

You do not take account of the fact that the data are proportions – where both the numerator and denominator are known – and are therefore limited to taking values between zero and one. The best approach for dealing with data of these form is to use generalised linear models (GLMs) (ref McCullagh and Nelder 1989[2]). This class of models extend the ideas of general linear models (simple linear regression) to analyse data that are proportions or counts and has all the advantages of linear models without having to make the assumption that the data and errors are normally distributed.

Difficulties that arise when you analyse the data in the way that you have because:

[2] McCullage P and Nelder JA (1989) *Generalized Linear Models*. Chapman and Hall. Second edition.

- You do not properly account for the fact that the total number of carcasses varies between sites and over time and so estimates of PIKE are not equally precise – i.e. the variance is not constant. For example, if 200 carcasses are found at a site and 100 of these were illegally killed this is a more precise estimate of PIKE than a site where 10 carcasses are found of which five were illegally killed. Thus observations from each site in each year are not equally precise. This does not seem to be accounted for in your modelling.[3]
- As you state in the paper, your model gives estimates and confidence intervals that are outside of the range zero to one. This is clearly incorrect.

The effect of these modelling errors is that the estimates of trend will be biased, although it is unpredictable in which direction this will go.

The strategy/argument used in the paper to justify the use of a simple linear model is ad-hoc, unreferenced and does not appear to be correct. Furthermore, the arguments that you present against the use of GLMs in Burn et al (2011) are wrong[4]. Specifically, you state that you use the simplest and most transparent model to deal with the data. But simple is not good if it is wrong. It is clearly preferable to have a more complex analyses that is correct. I'm afraid that we also did not find your approach at all transparent!

The paper refers to complex transformations used in nonlinear approaches (presumably our GLMs) and states that nonlinear approaches are "substantially more difficult to interpret, generally require stronger assumptions regarding the data generating process and do not allow all of the data to be used" (page 45). Later you state that to analyse data using nonlinear models you would need to exclude 32.1% of the data (page 47).

First, let's be clear that GLMs have a solid and robust theoretical foundation. There is a vast literature and knowledge in the statistical community (and in the wider research community) about the use of these methods. GLMs are actually quite intuitive, widely used and understood and not really all that complex and are taught on undergraduate statistics courses.

Second, the comment that they don't allow all of the data to be used is wrong. I assume that you are referring to those cases where all or none of the carcasses that were found were illegally killed – that is the proportion is zero or one. It is true that the logit transformation of zero or one is -/+ infinity. But in a GLM it is not the logit of the data that is modelled but the logit of the expected value of the data. Thus all the data can be used.

I note that despite your reservations about nonlinear models you do fit a Poisson regression to the number of carcasses that were illegally killed. I do not understand at all your reasoning for fitting a polynomial, or specifically a fifth order polynomial, for the number of legal carcasses.

Hierarchical data structure

The other key feature of the data is its hierarchical structure. By this I mean that there are repeated measurements taken from each site (you get a PIKE value each year) and in general there is data from more than one site within a country. You would expect measurements from the same site to be more

[3] Even if the numbers of carcasses were the same for each proportion, the variance is still not constant over the range of possible values, being greater near the middle. Generally, linear modelling is more robust to departures from normality than it is to non-constant variance.

[4] I am assuming that when you talk about nonlinear approaches you are referring to the use of GLMs because it was these, well an extension of these (see next section) that were used in other analyses of these data.

similar to each other than to other sites and measurements from sites within a country to be more similar to each other than sites from other countries. This needs to be taken into account – particularly when trying to account for the fact that sites are different to each other and that trends might differ between sites or countries.

In Burn et al (2011) multilevel or hierarchical modelling was used to account for this structure and the reasons for this approach given. Specifically, Bayesian hierarchical generalised linear models were fitted that also account for the fact that the data are proportions. This is a standard treatment of data with this kind of structure – for example see Gelman and Hill (2007)[5] – with a sound theoretical basis and great flexibility.

You have, to some extent acknowledged that the data have a hierarchical structure and the problems of ignoring this. You do not reference or refer to Burn et al (2011)'s approach for analysing these data which cope with the difficulties or state why you have chosen a different approach.

I am confused about your approach. It appears that you have only modelled a two level hierarchy (at country level) given your comment on page 12 that you have clustered your standard errors by country. You seem to have treated site as a fixed rather than random effect. I am unclear as to why you do this. Rather than repeat the main arguments here, I would suggest referring to Gelman and Hill (2007) for good descriptions of why the use of hierarchical modelling and random effects is a more sensible approach to the modelling than a fixed effects approach.

More general modelling observations

Although I see that you have explained why you get a difference in PIKE values for 2009 I am reminded from rereading Burn et al (2011) that we did carry out an analysis in which we omitted the sites where there was high natural mortality (which as you say leads to a decrease in PIKE even if illegal killing remains constant) and the decrease between 2008 and 2009 remained. This was to answer a query during the CITES CoP and was not published. Furthermore, both Burn et al (2011) and more recent analyses for example see https://cites.org/eng/prog/mike do not show a discontinuity from 2007 to 2008 but just a gradual increase from 2006/7 onwards.

Analysis of seizures data

As you know from Underwood et al (2013) the straight analysis of seizures data does not tell you much about the trade because of the biases in the seizure and reporting rates. Your analysis of the seizures data does therefore not match, in any way, the analysis presented in Underwood et al (2013) or other reports where we attempt to take account of these biases, correctly model the hierarchical structure and account for the fact that the data are counts. Furthermore, our analyses do not show a discontinuity in 2008 or 2009 and in reality the time delay that one might expect between illegal killing and illicit trade is likely to be variable and difficult to specify.

Fiona M Underwood & Robert W Burn
20 June 2016

[5] Gelman A, Hill J (2007) *Data analysis using regression and multilevel/hierarchical model*s. Cambridge University Press.