Notes to the talk

Thanks for the invite, I am very happy and honoured to be here, I love GEOMED and started attending back in 2009 in Charleston. So today I thought I could talk to you about the work we have been carried out around policy evaluation, in particular to show you how we thought of framing this in a hierarchical perspective, which obviously allows to account for space and time easily.

Talking about policy is a very broad topic – as they have the potential of impacting every aspect of life – I have here a few examples, starting with something that every person with a single bit of brain has hated badly – BREXIT, which has literally changed the UK (including affecting universities, but this is another story). Moving on, vaccinations (and I apologize for this picture which always provokes empathic reactions from the audience) is another one, which again is very broad and might depend on age and other characteristics and affect public health in general. Other examples are the sugar tax to tackle obesity, particularly in children, the smoking ban, which has appeared across several countries at different time point and then there are more environmental policies such as clean air zones implemented to reduce pollution and consequently to have a positive effect on health.

As these policies have the potential to influence every aspect of life, it is obviously imperative to evaluate if they work. They also usually bear a cost and that need to be accounted for. Unfortunately we find ourselves in a “post truth” era, where more often than not we get fake scientists providing opinion as if they were fact – which then become widespread and could potentially cause panic. So my point here being that it is of the utmost importance to use rigorous methods (basically rooted in statistics) to provide evidence based evaluation of policies.

In this context statisticians face a major problem which is that randomised clinical trials are not feasible – typically policy implementation does not involve random allocation of any sort – but there are quasi experimental designs which are generally a good idea- where the definition is here and as you see; there are two crucial aspects which are related to how to deal with dependencies in space/time and how to find comparable groups to make sure that we account for confounders

So here in this talk I will focus on the use of quasi-experimental designs in a longitudinal setting, with a focus on adding dependency. I will present three examples which comes from two different projects I was involved in the last few years and will finish with something which I find interesting and challenging (and which I have not tackled yet – which is spillover).

The first and probable most used design for longitudinal series is the interrupted time series design – which basically consider the time series of the outcome of interest which is then interrupted by the intervention or policy implemented. This plot shows the regression framework which the observed data and then specify an pre-intervention trend and a post intervention trend - there is the possibility of accounting for a change in level and a change in slope. What this does not include is controls and we cannot account for things like seasonality or more generally for external time varying effects.

If there are controls, another design which is generally very popular is DID – which has several ways of being specified – this is one, where the trends before-after are parallel and the only difference is a step change when the intervention takes place. There are alternative formulations where the trend changes after the policy starts for instance. Now this is used in a traditional causal inference perspective, which needs to abide specific assumptions – the parallel trends is the most important but then there is also the fact that the control allocations does not depend on the outcome and the composition of exposed and control groups are stable across time. This means that in this formulation this cannot be used if the intervention in rolled out for instance. An additional issue here as well is how to account for dependence which probably somehow would violate the assumption of control/exposed allocation. In any case this is a popular design. There are similar designs called regression discontinuity (a quasi-experimental impact evaluation method used to evaluate programs that have a cutoff point determining who is eligible to participate), segmented regression (a method in regression analysis in which the independent variable is partitioned into intervals and a separate line segment is fit to each interval.)

These methods have things in common, they have a before and a after, which you can things of specify through regression lines (or something more complex if appropriate). What I want to do hrre is try to find a specification which is common and more general, so that we can then fit ITS, DID, etc. but also overcome some of the limitations (like that the interventions need to start at the same point for everybody or that the treatment status of one unit influence the others). In particular I am going to sue the Bayesian hierarchical framework for this.

So I want for each unit (individual or area) and each time an outcome y\_it, an indicator variable for the exposed (meaning those you get the intervention w\_i and an indicator variable for twhen the intervention starts z\_t; for the moment I keep z only indexed by t assuming that all the units get the intervention t the same time and w only indexed by i meaning that each unit either get or not the intervention. Then we get a linear predictor as a baseline level of the outcome in the two groups a linear trend and a step change in the intervention – this is a DID, then if we get the lambda\_t which is bold as it can include a trend before and a trend after the intervention we then get an ITS (considering that w\_i is 1 everywhere so in that case the effect for the step change are the sum of delta0 and delta 1). Why hierarchical, well – it is undoubtedly flexible, we can include controls and potential non linearity and dependency for instance through a gamma\_i; why Bayesian, well it allows us to regularize inference still using vague priors if for instance we consider penalized complexity priors – ASK GIANLU Additionally it allows to give direct estimates of the uncertainty – it is easy to rescale at whatever resolution we want, which is particularly helpful when we have generalized linear models and want to obtains RR or scores or more generally functions of the parameters.

Now I would like to focus on some examples as a way of showing you how we used this framework to actually evaluate policy implementation. So the first example stems from a EU policy back in 2010 which wanted to limit the amount of waste sent to landfills so pushed for the constructions of incinerators. This created a tension as on one hand the EU regulation also said something about how the incinerators needed to be built – the fact that administrations were embracing this to reduce waste, and the local population which perceived this as potentially dangerous (see below) fuelled by the lack of studies showing evidence of an effect or the lack of an effect. So in all this PHE and the Scottish government funded a study to investigate the effect of incinerators on birth outcomes. This was a huge study, which involved different analyses, from an emission perspective as well as several epidemiological analyses on birth outcomes and on breast milk quality. Here I will focus on the first paper which wanted to use natural experiments to see if there was any signal of an effect of the building of the opening of new incinerators on infant mortality.

We had 8 MWI opening within 2003-2010; exposed areas were MSOA with a centroid within 10km – we also needed to account for trends in the outcome and so we considered controls – the choice was based on a series of characteristics related to socio-demographics as well as population at risk (live births). We actually needed a bit of a complex system to select the control areas which is depicted here – starting with the area of interest, we got all the MSOA within the same region and built 10km buffers making sure to remove the ones where a bit was within the study area. Then using a distance we found the closest in terms of the variables indicated here and used that as control area. On the right you see the incinerators and the respective control areas – POLICY IMPLEMENTED AT DIFFERENT TIMES!

Then the model is here and is based on the framework I presented before – we have the baseline for the areas under the incinerators and for the controls, a temporal trend before for both and a temporal trend after. On the spatial random effect we include a combination of structured and unstructured terms. We have the additional issue of dealing with disconnected components, which we can account for using a scaling of the precision, I am not going to go into details here but just to mention that this is also something we can deal with easily within this framework.

Here I have recapped the linear predictors for each group, so that it is easier to see how we can interpret each term – basically if on the column we have the intervention (opening of incinerator) a nd on the rows we have the exposed/controlled regions, you can see that we can compare several things – psi1 gives the difference in the trends before the intervention, which we would like to be around 0 to make sure the trends are parallel before the intervention, delta0 gives us the difference before-after for the controls, which if different from 0 might indicate some temporal confounding; then delta1 gives the difference in the before/after trends in exposed vs controls. Note that we have all the linear predictors, so a rich output so we can look at differences before/after at specific time points and for specific areas.

Looking at the results we did not find any difference after the opening of the incinerators comparing the exposed areas and the controls. As also looked at sex ratios and while there was a slight difference, there was too much uncertainty. SAY SOMETHING ABOUT THE STUDY OF BECKY

I am now move on to a different example, which is part of a project funded by the Wellcome trust. We should assume that one of the core function of any government is to promote the wellbeing of all its citizens, in an equitable way. This would be the role of a welfare state and it should be obtained through a set of policies, rather broad, which encompass migration, taxation, welfare reform, public health care. To set the scene remember the economic recession in 2007-2008; soon after that UK had a new government which started introducing austerity policies . At the same time mental health of the UK population (and particularly of some vulnerable groups) seemed to have deteriorated over the years. What we wanted to do was to estimate if there is any evidence that specific policies have had a bad effect on mental health. This is a broad project which has been funded by the Wellcome Trust under the collaborative awards in science. So here I want to talk to you about two examples: the first is around universal credits.

UC are a policy which was introduced by the coalition government which was actually a Tory government in theory to simplify benefits (**Housing Benefit. income-related** Employment and Support Allowance (ESA)). Maybe their heart was in the right place, I kind of doubt it but that is something I can’t really have hard evidence of, but what happened was that the change was not smooth and resulted in lengthy delays for people who were already fragile and should have been helped.

Here we used the UK household longitudinal study and focused on people of working age (16-64) with employment status, residence and who responded to at least one mental health questionnaire. We used the GHQ which is a standardized qwuestionnaire to evaluate the level of anxiety and depression. Before I go into the model, which is similar to the previous one, I need to focus on the definition of exposure vs controls – this is tricky and a little bit fuzzy, but we used unemployment status, as in the vast majority UC are given to unemployed. Hence employed are our control – we removed students and people with longlasting health conditions which would qualify for a different type of benefit. Another issue here is the policy definition, as unfortunately we do not have individual information about UC. So what we needed to do was to use areas as a proxy – basically we used a definition of contextual awareness, we know the % of people in each area (MSOA) which start receiving UC and we assume that when an area reaches 25% that area become contextually aware, meaning that even if someone living there does not receive the credits they know it is going to happen soon and might become anxious. We also accounted for individual level characteristics such as sex age etc. I have a similar model as before, but with a slight additional complication which is that for each local authority I have a different time when UC start so I have included here t\_i. Even more we are assuming here that there might be a sudden change in the outcome around the time when the policy is implemented, so I include also that into the model. We also have random effects which include strata and area level effects. TEMPORAL RANDOM EFFECT?year rather than months?

Here I have included the effects – and you can see that for the controls we do not see any evidence of a change, while if we look at the difference between exposed and controls we see a very large step change (delta0\*), while then the sustained effect is negative, which means that over time the level seems to go back down. HOW DO I INTERPRET

An additional way of presenting the results relies of calculating a score which could synthesise the effect of the policy - we do that through a standardize % change – which gives us the probability (TO CHANGE) of experiencing the outcome before vs after after adjusting for what happens in the controls – we get the distribution of the average linear predictor after the policy for the exposed and then we adjust the one before for the ratio between the after/before in the controls, marginalized for the confounders (check how, using the average?) – this means that if the ratio is above 0 it will inflate the before mean so will reduce the impact of the policy as presumably there is some temporal residual confounding. What we get is here, on the left it shows how in space it varies, with changes going from -50% to 100% - on the right we also see the full extend of the uncertainty. Note that the national change is a bit less than 3% in the probability of experiencing mental distress.

The last example is still within the same project but looks at a specific group, which is the Windrush generation. If you are not familiar with this, here is a bit of a story which goes back to after the second world war, when in the UK there was a lack of working resources – so there was a massive advertisement of the beauty of England and how easy was to find job to the colonies, in particular the Caribbean – the results is that lots of people moved from there – the incentive was that they were offered indefinite leave to remain (not citizenship!). Then an immigration act in 1971 limited the nb of people who could stay. In 2020 the home office moves and the result is that somehow they misplaced all the paperwork (still on paper!) of all those immigrants – this meant that then people were liable to become illegal. To make things worse the hostile environment policy in 2012 and immigration act in 2014 made things like getting healthcare, rents etc difficult. This culimnated in 2017 with a scandal which went widespread on the newspapers and costed the job to the home secretary.

What we wanted to do here was to evaluate if there was evidence of a negative impact on mental health of the policies and of the scandal, specifically focusing on ethnic minorities. We used again the Understanding society data and the GHQ score, but here we have 2 policies and also multiple exposure groups as we wanted yo separate across ethnic minorities. The study design is depicted on the left identifying the 3 periods and the corresponding waves of the longitudinal study. The model, which by now you are I am sure loving is on the right and I have put in red the differences between this and the previous one – we have difference indexes for w and for z resulting in more parameters. We still account for an immediate effect and a sustained effect

What we find is here in terms of difference in the change of GHQ score between each group and the reference one which is white. What we can see is that the immediate effect is evident across black Caribbean, African, but also Indian and Pakistani, which might be an effect of a stressful situation and the fact of being perceived minorities and migrants – it seems that the news scandal has even a greater effect. The sustained effect is less clear for black Caribbean, but seems to be stronger for black African.

The final things I wanted to mention is something which can mess things up and so we need to work on this. It is mostly related to policies which are more area based – for instance the ULEZ. You see here on the right

Can we really infer causality – well I believe we go as close as possible – conditional on the fact that we get good controls and we adjust for all the potential confounders. Is this the case in reality? I don’t know, I believe we can try our best. If you want to place yourself in the context on classic causality there is a set of assumptions which needs to be met, such as trends which are parallel, stable groups, no spillover. Some of these hold, some don’t. But as this is a completely model based approach, what I would say is that in this context we can easily test these assumptions and this would also tell us something about how much we can interpret this in a classic causality context. If you want to go as close as possible to potential outcomes we can use posterior predictive distributions to simulate the counterfactual scenarios.