Portfolio

A look at my most recent projects.

Overview

1 **How States Really Respond to Status Dissatisfaction: a closer look at the material and temporal dynamics of status-driven conflict**
A research paper based on analysis in Stata View project

2 **Data Analytics at Costcutter Canterbury**
An on-going business initiative utilising SQL, Power BI and R View project

3 **A Method to Enable Time Series Analysis of Multiple Deprivation Indices with an Illustrative Analysis of Mortality Rates**
A research paper based on analysis in R View project

1 How States Really Respond to Status Dissatisfaction: a closer look at the material and temporal dynamics of status-driven conflict

A research paper based on analysis in Stata

1.1 Project summary

Aims
To demonstrate that the effect of status-dissatisfaction on the probability of a country initiating a militarized conflict is highly dependent on *a*) its material capabilities and *b*) its past experiences of war.

Methods
Applying advanced modelling techniques to build robust logistic regression models using time series panel data. The analysis is structured as follows:
1) I use cubic spline transformations to account for hidden non-linearity in key control variable, highlighting the effects of misspecification on the substantive interpretation of the main effect - and it's dependence on the control effect - via diagnostic/effect plots.
2) I employ experimental Random Effects Within Between models to isolate and plot time-dependencies in the main effect while simultaneously accounting for dependence on the control effect via cross-level three-way interaction terms.

Outcomes
I substantiate the above hypotheses via statistical analysis and graphical visualization, drawing original conclusions that call into question the effectiveness of the conflict management strategy of appeasement.

Skills
Building sophisticated statistical models to test theory-driven a priori hypotheses.
Cleaning, reshaping and combining many datasets from different sources.
Use of state-of-the-art panel data techniques.
Understanding of the merits and limitations of basic and advanced statistical concepts/methods.
Creating informative and highly accessible data visualisations.
Communication of highly abstract ideas/complex methods in a manner accessible to both statistical and non-statistical audiences.
Stata coding of all analysis across 30+ scripts/do-files.

Context

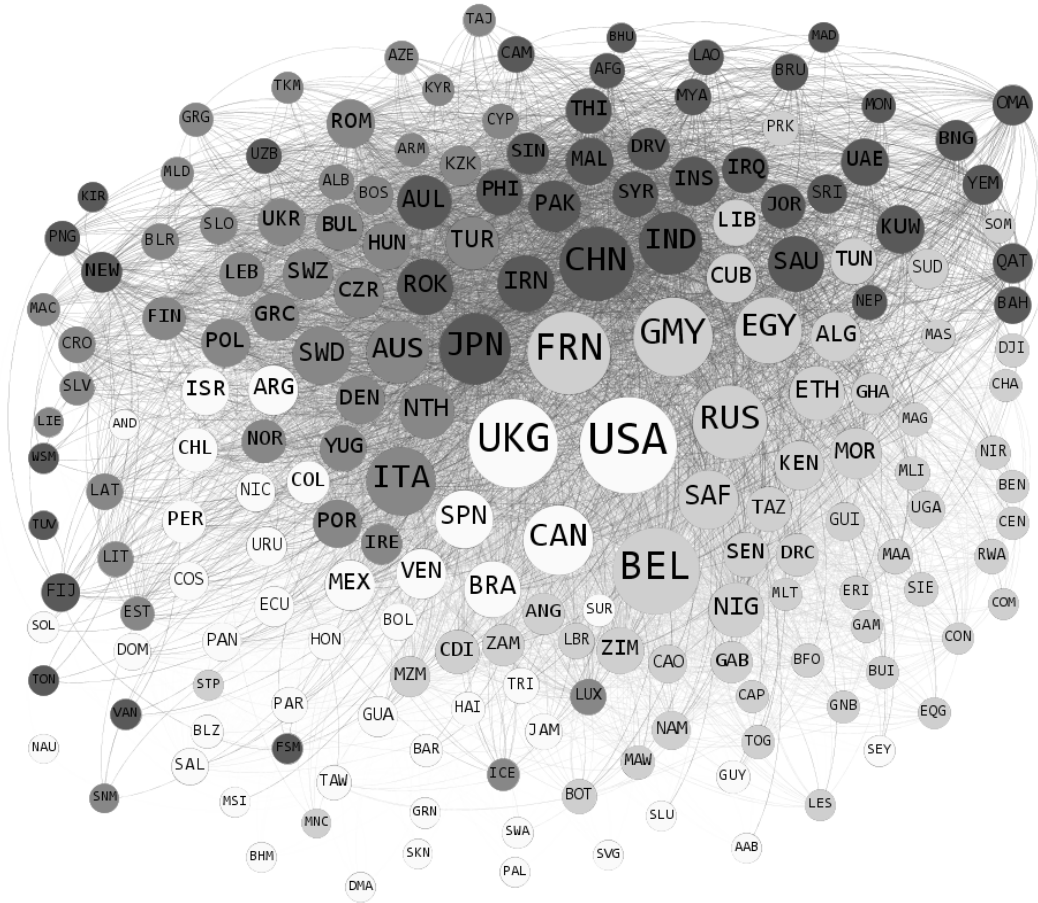
Data structure
Panel unit: pairs of countries.
Time unit: every year between 1949 and 2000 (contingent on data availability).
N: 950,000 +.

Predictor of interest
Status deficit: a measure of the degree to which country *i* is dissatisfied with its social status.

Outcome variable
MID initiation: binary variable where 1 denotes the initiation of a militarized dispute by country *i* against country *j*.

Moderator variables
CINC: a composite measure of country *i*'s material capability.
Peace years: number of years since country *i* last initiated an MID.

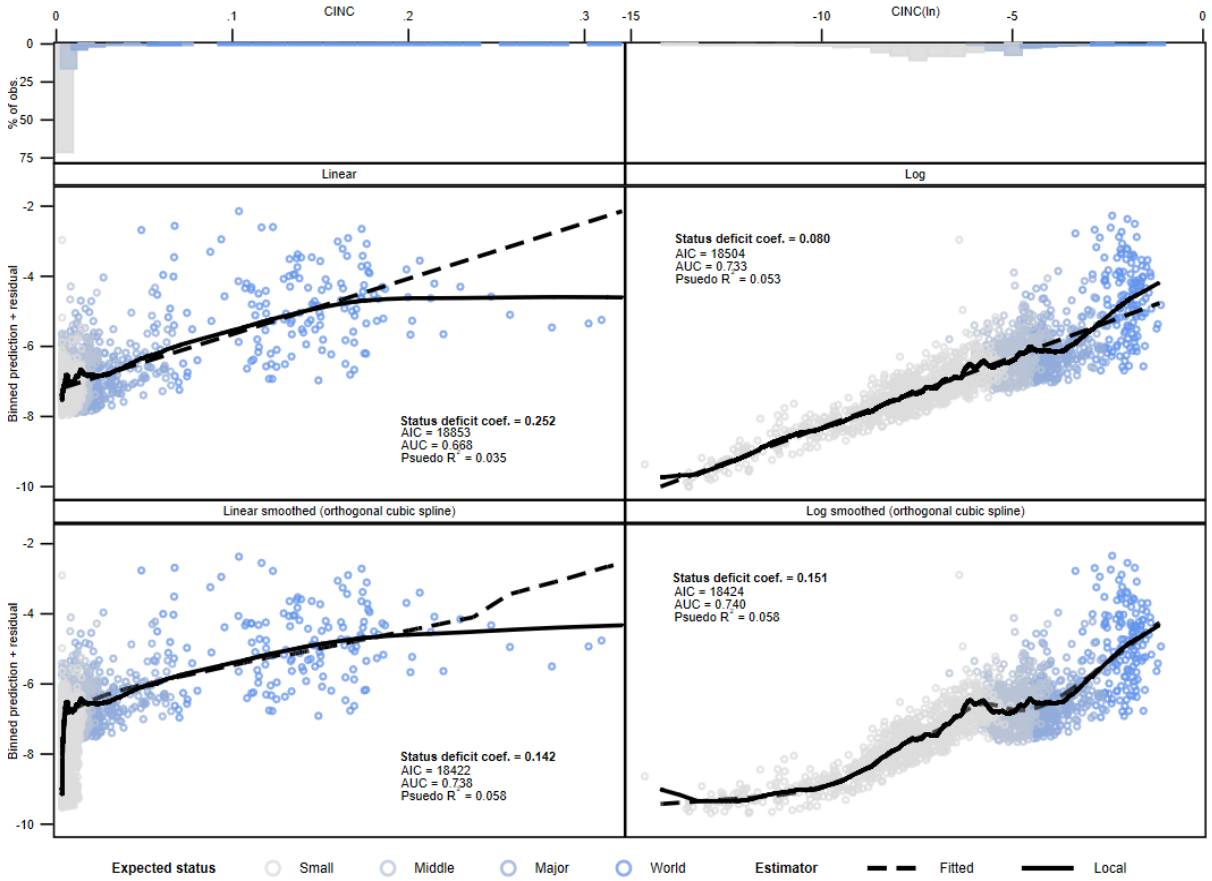
Figure 1.2.1: *The inter-state system in 2000 by status rank (node size) and derived community membership (color)*



Skills 1.2.1
Reshaping data for Social Network Analysis in Gephi.
Creating dynamic variable using PageRank and community detection algorithms.
Multi-level visualization of complex social networks.

Modelling nonlinearity in capabilities moderator variable

Figure 1.3.1: *Component-plus-residual plot of linear and cubic spline functions of CINC versus CINC(ln)*



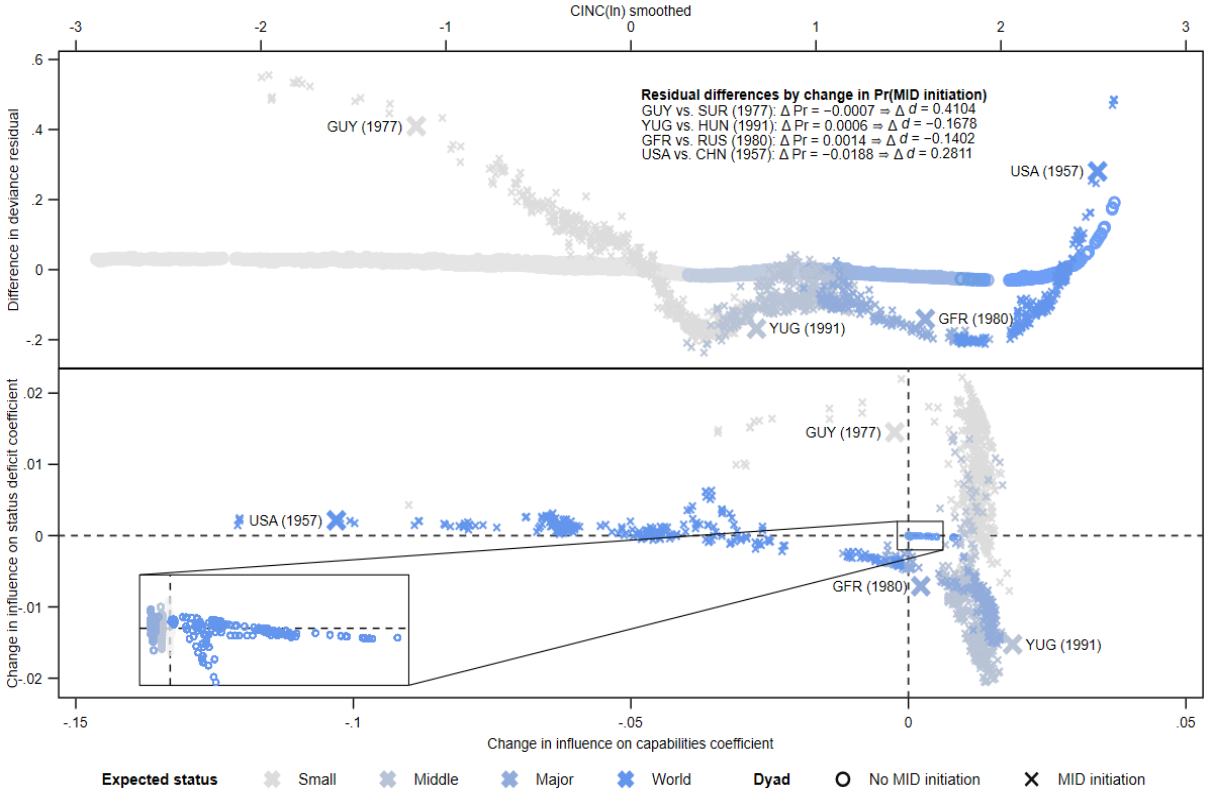
Note. Component-plus-residual (CPR) plot of the estimated effects of linear and cubic spline functions of CINC and CINC(ln) on the probability of initiating an MID. Histograms of CINC in its raw form and the natural log of CINC are plotted in the top-left and top-right panels respectively, while CPR estimates of CINC/CINC(ln) as *a*) a linear function and *b*) an orthogonalised cubic spline of 7 percentalised knots are plotted in the corresponding middle and bottom panels. Estimates are the sum of the partial linear prediction and deviance residual of a given observation, fixing status deficit at the full population mean while allowing CINC/CINC(ln) to vary. Estimates are averaged within (roughly) equally sized bins of CINC/CINC(ln). Each binned estimate is represented by a circular marker, gradating in color across levels of expected status from small states (dark) to world powers (light). Dashed green lines represent best fit rom OLS regression of CPR estimates on CINC/CINC(ln). Solid purple lines represent locally smoothed fit from LOWESS regression using a bandwidth of .05. The coefficient for status deficit from each model is reported in bold, with goodness-of-fit statistics below.

Stata code 1.3.1: *Binning of component-plus-residual estimates by sub-population*

```
*** Loop over levels of expected status (subpopulations)
local j 0
local sample small_1 middle_1 major_1 world_1
qui foreach pc in `sample' {
    local ++ j
    *| Temp variable names
    tempname obs`j' binno`j'
    sort mcap_1_lg1_ln
    *| Observation number by subpopulation
    gen `obs`j'' = _n if pr_d < . & `pc' == 1
    su `obs`j'', meanonly
    local subobs = r(N)
    *| Number of bins (proportional to subpopulation size)
    local nbins`j' = floor(sqrt(`subobs'))
    local nbins`j': di %3.0f `nbins`j''
    *| Bins of (roughly) equal size
    egen `binno`j'' = cut(`obs`j'') if pr_d < . & `pc' == 1 , group(`nbins`j'') icodes
    *| Average CINC value by bin
    egen av_d`j' = mean(mcap_1_lg1_ln), by(`binno`j'')
    *| Average prediction by bin
    egen avcpr_d`j' = mean(pres_d) if `pc' == 1, by(`binno`j'')
    *| Tag binned observations
    egen tag_d`j' = tag(`binno`j'')
}
```

Skills 1.3.1
Understanding of the substantive effects of variable misspecifiction.
Using advanced variable transformation techniques to maximize explanatory power.

Figure 1.3.2: *Diagnostic plot for models with CINC versus cubic spline of CINC(ln)*



Note. Plot of the difference in *a*) residuals and *b*) influence for models of status deficit plus CINC versus CINC(ln) smoothed. Upper panel plots the difference in deviance residual against standardized values of CINC(ln) smoothed. Residual differences, and corresponding change in predicted probability, are reported for four example dyads - one for each expected status rank. Lower panel plots change in positive influence on coefficients for status deficit and capabilities from models with CINC versus CINC(ln) smoothed. Estimates are the difference in standardized DFEBTA(S). More specifically, the difference in *a*) the DFBETAS of status deficit and *b*) the DFBETAS of CINC and the cross-spline mean of the DFBETAS of CINC(ln) smoothed. Both upper and lower panels plot markers for each observation in the multivariate sample. Markers gradate in color across levels of expected status, with crosses and circles representing the initiation and non-initation of an MID respectively. A 'zoomed-in' inset plot of DFBETAS estimates for non-initiating dyads is included in the lower panel as an additional visual aid.

Stata code 1.3.2: *"Zoomed-in" inset-plot on invisible 2nd axis*

```
*** Add inset plot axes and connecting lines
addplot 2: (scatteri -.002 -.002 .002 -.002 .002 .00625 -.002 .00625 -.002 -.002, recast(line)
     lpattern(solid) lcolor(black)) /* Superimposed inset box
 */ (pcarrowi -.002 .00625 -.021 -.09, recast(line) lpattern(solid) lwidth(vthin) lcolor(black)) /*
      Connecting lines
*/  (pci .002 -.002 -.0055 -.1385, lpattern(solid) lwidth(vthin) lcolor(black)) (scatteri -.021 -.09 -.021
     -.1385 -.0055 -.1385 -.0055 -.09 -.021 -.09 , recast(line)
    lpattern(solid) lcolor(black)) /* Inset plot along hidden 2nd X-axis
*/  (pci -.021 -.133 -.0055 -.133, lpattern(shortdash) lcolor(black) lwidth(thin)) (pci -.013 -.1385 -.013
     -.09, lpattern(shortdash) lcolor(black) lwidth(thin) norescaling
    legend(off)) /* Axis guide lines */

*** Width of axis guide lines
gr_edit .plotregion1.graph2.plotregion1.plot19.style.editstyle line(width(vthin)) editcopy
gr_edit .plotregion1.graph2.plotregion1.plot22.style.editstyle line(width(vthin)) editcopy

*** Plot inset markers on 2nd X-axis
addplot 2: (scatter DFZdif_def DFZdif_mcap if pclass_1 == 1 & DFZdif_mcap <=.002, msymbol(oh) msize(small)
     mcolor(gs1%2) xaxis(2) yaxis(2) ylabel(-.0001 .000975, axis(2)) /* Small powers
*/  xlabel(-.0007 .00725, axis(2)) yscale(axis(2) off) mlwidth(thin) jitter(.1)) (scatter DFZdif_def
     DFZdif_mcap if pclass_1 == 2 & DFZdif_mcap <=.002, msymbol(oh) msize(small) /*
*/  mcolor(gs5%25*1.2) xaxis(2) yaxis(2) ylabel(-.0001 .000975, axis(2)) xlabel(-.0007 .00725, axis(2))
     yscale(axis(2) off) mlwidth(thin) jitter(.1)) /* Middle powers
*/  (scatter DFZdif_def DFZdif_mcap if pclass_1 == 3 & DFZdif_mcap <=.002, msymbol(oh) msize(small)
     mcolor(gs9%25*1.1) xaxis(2) yaxis(2) ylabel(-.0001 .000975, axis(2)) /* Major powers
*/  xlabel(-.0007 .00725, axis(2)) yscale(axis(2) off) mlwidth(thin) jitter(.1)) (scatter DFZdif_def
     DFZdif_mcap if pclass_1 == 4 & DFZdif_mcap <=.002, msymbol(oh) msize(small) /* World powers
*/  mcolor(gs13%50) xaxis(2) yaxis(2) ylabel(-.0001 .000975, axis(2)) xlabel(-.0007 .00725, axis(2))
     yscale(axis(2) off) xscale(axis(2)off) mlwidth(thin) norescaling legend(off))
```

Skills 1.3.2

Implementation/illustration of comprehensive and innovative diagnostic methods.
In-text signposting plus inset plot to enhance readability.

Table 1.3.1: *Table of coefficients for models with CINC versus cubic spline of CINC(ln)*

	With CINC		With CINC(ln) smoothed	
	(1)	(2)	(3)	(4)
Main effects				
Status deficit	0.252**	0.248**	0.151**	−0.0386
	(0.0479)	(0.0474)	(0.0525)	(0.0691)
CINC	0.312**	0.300**	-	-
	(0.0195)	(0.0205)		
CINC(ln) smoothed				
k0	-	-	0.943**	1.080**
			(0.109)	(0.134)
k1	-	-	−0.0215	0.158
			(0.0911)	(0.123)
k2	-	-	0.154	0.0430
			(0.0855)	(0.0886)
k3	-	-	−0.0459	−0.0244
			(0.102)	(0.0966)
k4	-	-	−0.0655	−0.0849
			(0.0673)	(0.0684)
k5	-	-	−0.118*	−0.117*
			(0.0539)	(0.0532)
k6	-	-	−0.151**	−0.156**
			(0.0442)	(0.0443)
k7	-	-	0.0354	0.0457
			(0.0368)	(0.0414)
Constant	−6.776**	−6.783**	−7.092**	−7.156**
	(0.0627)	(0.0630)	(0.0814)	(0.0884)
Interactions				
Status deficit X CINC	-	0.0760	-	-
		(0.0481)		
Status deficit X CINC(ln) smoothed				
k0	-	-	-	0.400**
				(0.119)
k1	-	-	-	0.379**
				(0.108)
k2	-	-	-	−0.247**
				(0.0653)
k3	-	-	-	0.124
				(0.0724)
k4	-	-	-	0.0505
				(0.0508)
k5	-	-	-	−0.0351
				(0.0389)
k6	-	-	-	−0.0279
				(0.0433)
k7	-	-	-	0.000638
				(0.0372)
Observed	965,624	965,624	965,624	965,624
Psuedo R^2	0.0352	0.0355	0.0578	0.0590

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$

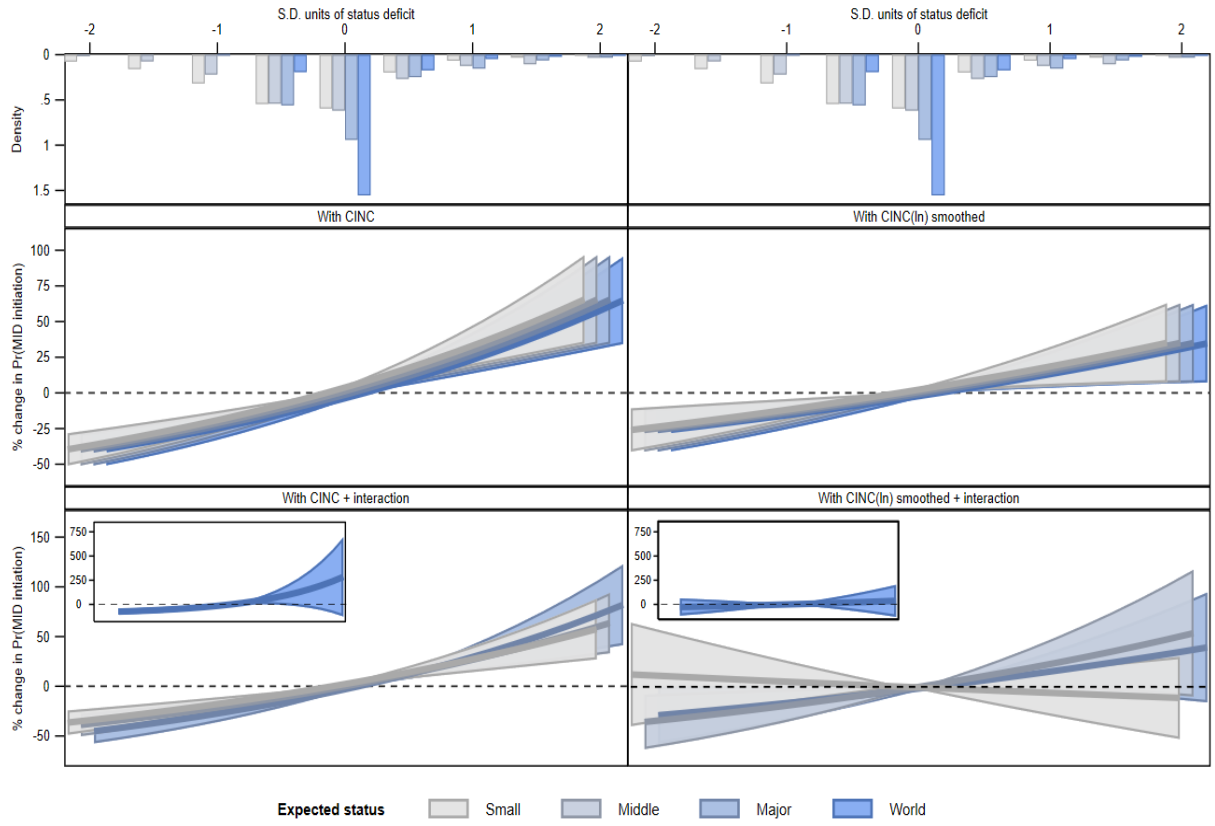
LaTeX***code 1.3.2:*** *Partial construction of coefficient table*

```
% Table environment
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
begin{tab}[!htp] % Set custom floating environment
    \caption{\raggedright{Table of coefficients for models with CINC versus cubic spline of CINC(ln)}} %
        Caption
    \tiny % Font size
    \noindent\makebox[\textwidth]{ % Table dimensions
    \begin{tabular}{@{}l*{4}{D{.}{.}{-1}}@{}} % Cell alignment by decimal point
        \toprule \toprule &\multicolumn{2}{c}{\textbf{With CINC}} & \multicolumn{2}{c}{\textbf{With
            CINC(ln) smoothed}}\\ % Set multi-column structure
        \cmidrule(lr){2-3}\cmidrule(lr){4-5} &\multicolumn{1}{c}{(1)} &\multicolumn{1}{c}{(2)}
            &\multicolumn{1}{c}{(3)} &\multicolumn{1}{c}{(4)} % Ruling
        \\ \midrule  % More ruling
```

```
% Main effect coefficients + SE values
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
\\[-2.75mm] \textbf{\emph{Main effects}} \\ \addlinespace % Main effect cell heading
\hspace{5mm}Status deficit &0.252\sym{**}  &0.248\sym{**}  &0.151\sym{**} &-0.0386 \\ &(0.0479)
     &(0.0474)  &(0.0525)  &(0.0691) \\
\addlinespace
\hspace{5mm}CINC &0.312\sym{**}  &0.300\sym{**}
     &\multicolumn{1}{r}{-}{\RaggedRight}\hspace{5mm}&\multicolumn{1}{r}{-}{\RaggedRight}\hspace{6.5mm}
     \\ &(0.0195)  &(0.0205)  &  &  \\
```

Skills 1.3.1(b)
Employing advanced table creation techniques in LaTeX
Using conventional coefficient/significance signs to complement substantive effect plot (see below).

Figure 1.3.3: *Plot of the substantive effect of status deficit for models with CINC versus cubic spline of CINC(ln)*



Note. Plot of the substantive effect of status deficit on the probability of initiating an MID for models with CINC versus CINC(ln) smoothed. Estimates represent relative risk (RR) from the mean, fixing CINC/CINC(ln) smoothed at the expected status mean while switching status deficit from -2 to 2 in increments of .2 S.D.. Estimates for models with CINC and CINC(ln) smoothed are plotted in the left and right panels respectively, while corresponding middle/bottom panels plot estimates from models which omit/include an interaction term between status deficit and CINC/CINC(ln) smoothed. Both top panels plot histograms of status deficit, based on probability density estimates within levels of expected status. Mean RR estimates are represented by solid lines and bounded by confidence interval areas at the 95% level. Estimates gradate in color across levels of expected status. A 'zoomed-out' inset plot of RR estimates for world powers is included in both bottom panels so as to prevent visual compression.

Stata code 1.3.3: *Calculation of % change estimates by sub-population*

```
*** Loop over expected status (sub-populations)
local j 0
qui foreach pc of varlist small_1 middle_1 major_1 world_1 {
     frame copy rpc rpchange_`pc'    // New frame
     frame change rpchange_`pc'
     local ++ j *| // Set increment by sub-pop
     local j2 0
     *| Loop over models
     forvalues i = 1/4 {
         local ++ j2    // Set increment by model
         est restore pr`i'    // Restore model estimates
         mat rpc`i'_`pc' = J(21,3,.)    // Matrix to store results
         *| Predictions for non-spline models
         if `j2' == 1 | `j2' == 2 {
             margins, subpop(if `pc' == 1) vce(unconditional) at(defz=(-2 (.2) 2)) atmeans post
             est store mar`i'_`pc'
         }
         *| Predictions for spline models
```

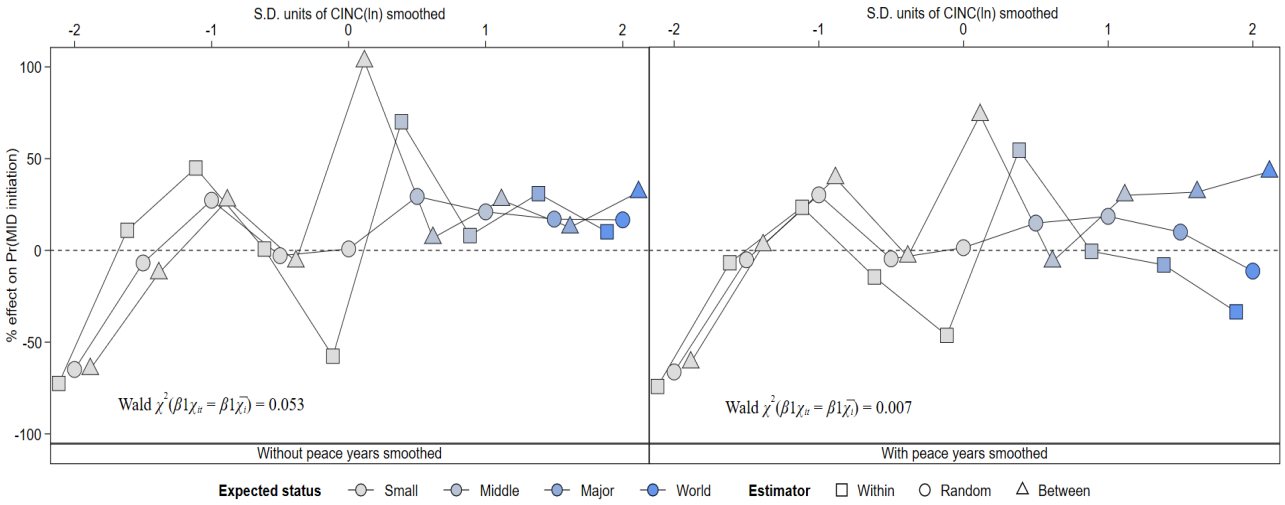
```
else {
    tempvar avmcaplnk_0
    tempvar absmcaplnk_0
    egen `avmcaplnk_0' = mean(mcaplnk_0) if `pc' == 1    // Sub-population mean of CINC/CINC(ln)

    gen `absmcaplnk_0' = abs(`avmcaplnk_0' - mcaplnk_0)    // Temp var to hold sub-pop mean
    sort `absmcaplnk_0'
    forvalues k = 0/7 {
        local k`k'av = mcaplnk_`k'[1]    // Macro for knot value at sub-pop mean
    }
    margins, subpop(if `pc' == 1) vce(unconditional) at(defz=(-2 (.2) 2) mcaplnk_0 = `k0av'
         mcaplnk_1 = `k1av'/*
    */ mcaplnk_2 = `k2av' mcaplnk_3 = `k3av' mcaplnk_4 = `k4av' mcaplnk_5 = `k5av' mcaplnk_6 =
         `k6av' mcaplnk_7 = `k7av') post
    qui est store mar`i'_`pc'
}
*| Loop over margins estimates in .2 sd increnents of status deficit
forvalues k = 1/21 {
    if `k' != 11 {
        if `k' == 1 {    // 1st est. (due to different var suffix)
            nlcom (rpc`k':(_b[`k'bn._at]/_b[11._at]-1)*100), post    // Calculate % change
            qui mat rpc`i'_`pc'[`k', 1] = e(b)    // Store est.
            qui mat rpc`i'_`pc'[`k', 2] = e(b) - invnorm(.975) * _se[rpc`k']    // Store lbCI
            qui mat rpc`i'_`pc'[`k', 3] = e(b) + invnorm(.975) * _se[rpc`k']    // Store ubCI
            qui est restore mar`i'_`pc'
        }
        else {    // Remaining est. (except where status deficit = 0)
            nlcom (rpc`k':(_b[`k'._at]/_b[11._at]-1)*100), post    // Calculate % change
```

Skills 1.3.3
Graphing main effect in highly accessible % terms.
Understanding of the effects of omitting/misspecifying interaction terms.

1.4 Modelling time-dependence in status deficit main effect

Figure 1.4.1: *Plot of the within, random and between effects of status deficit for models which exclude versus include a cubic spline of peace years*



Note. Plot of the within, random and between effects of status deficit on the probability of initiating an MID for models which exclude versus include a smoothed function of peace years. Estimates represent discrete percentage effect from the mean, fixing CINC(ln) smoothed at incremental values of .5 S.D. while switching status deficit from 0 to 1 (and allowing peace years smoothed to vary). Random effect estimates - represented by circular markers - are drawn from standard RE models with status deficit, CINC(ln) smoothed and an interaction term between these effects. Estimates of the within and between effects - represented by square and triangular markers respectively - are drawn from REWB models with separate within and between estimates for status deficit and CINC(ln) smoothed - plus a cross-level interaction between the within-effect of status deficit and the between-effect of CINC(ln) smoothed. Estimates from models which include peace years smoothed - an orthogonalised cubic spline of 4 knots - are plotted in the right-side panel. Both RE and REWB models estimate the random effect of peace years smoothed. *p*-values from a Wald χ^2 test of equality of REWB status deficit coefficients are reported in each panel.

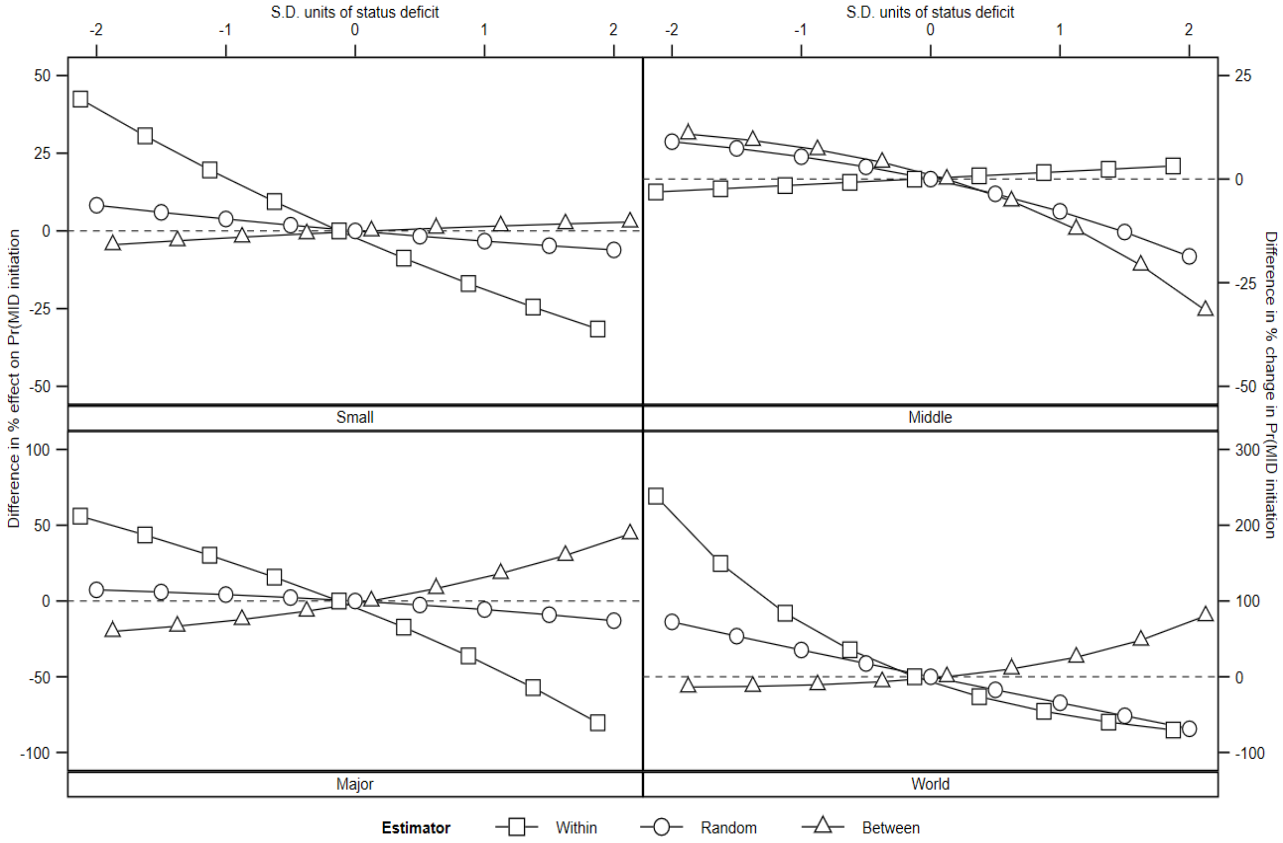
Stata code 1.4.1: *Combining graphs plus Wald test text signposting*

```
*** Combine graphs
grc1leg2 not t, rows(1) ysize(3) xsize(7.5) imargin(l=-4 r=-2.25) graphregion(margin(l=3.5 b=0 t=-2.5))
     legendfrom(not)
*| Add Wald chi-squared statistics
local wald1: di %4.3f scalar(wald1)
local wald2: di %4.3f scalar(wald2)
```

```
addplot 1:(scatter _e _at if tag == 10, text(-83.5 -1 "{stSerif:Wald
    {it:{&chi}}{super:2}({it:{&beta}}1{it:{&chi}{subscript:it}} = {it:{&beta}}1{it:{&chi}} /*
*/ `=ustrunescape("\u0305")'{it:{sub:i}}) = `wald1'}", size(medlarge))), norescaling legend(off) /*
*/ addplot 2:(scatter _e _at if tag == 10, text(-85 -1 "{stSerif:Wald
    {it:{&chi}}{super:2}({it:{&beta}}1{it:{&chi}{subscript:it}} = {it:{&beta}}1{it:{&chi}} /*
*/ `=ustrunescape("\u0305")'{it:{sub:i}}) = `wald2'}", size(medlarge))), norescaling legend(off)
```

Skills 1.4.1
Understanding/visualization of the limitations of standard panel data methods
Implementation of experimental Random Effects Within Between models

Figure 1.4.2: *Plot of the difference in the within, random and between effect of status deficit for models which exclude versus include a cubic spline of peace years*



Note. Plot of the difference in the within, random and between effects of status deficit on the probability of initiating an MID for models which exclude versus include peace years smoothed. Estimates are the difference in discrete percentage effect from the mean, fixing CINC(ln) smoothed at the expected status mean while switching status deficit 0 to 1 (and allowing peace years smoothed to vary). Random effect estimates - represented by circular markers - are drawn from standard RE models with status deficit, CINC(ln) smoothed and an interaction term between these effects. Estimates of the within and between effects - represented by square and triangular markers respectively - are drawn from REWB models with separate within and between estimates for status deficit and CINC(ln) smoothed - plus a cross-level interaction between the within-effect of status deficit and the between-effect of CINC(ln) smoothed.

Stata code 1.4.2: *RE and REWB Generalised Structural Equation Models to enable calculation of difference in % change*

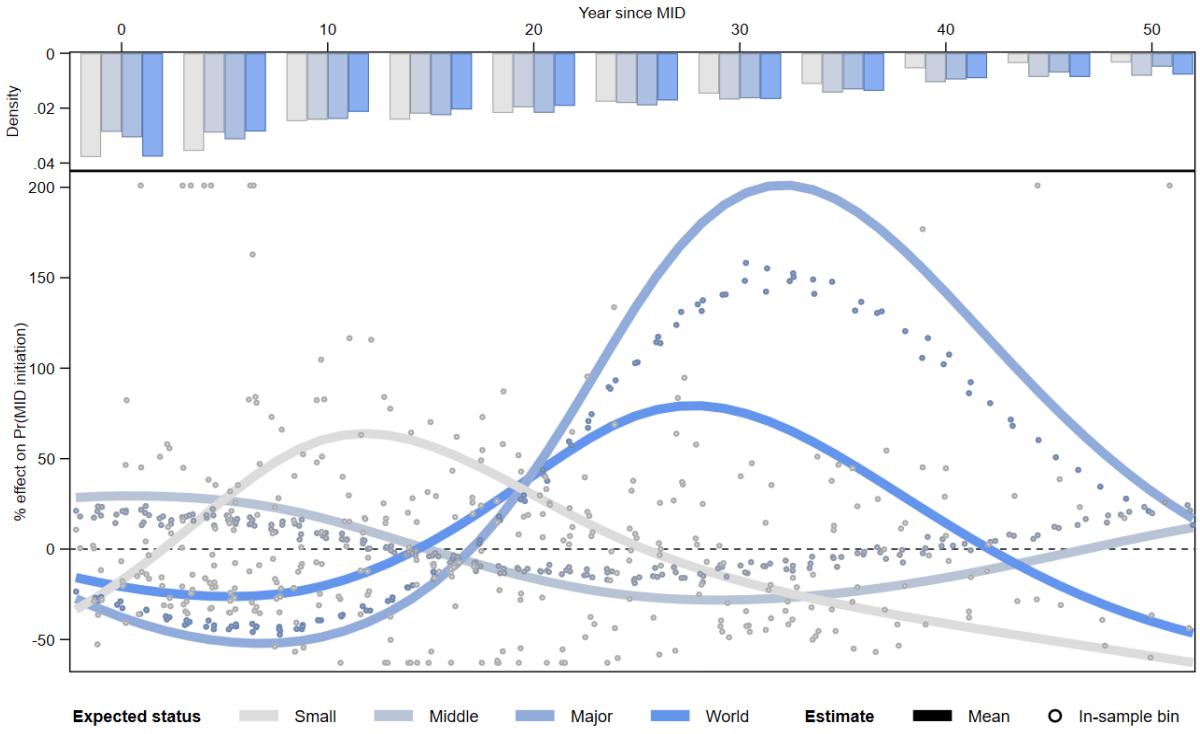
```
*** Random effects GSEM
clonevar midint1 = midint
clonevar midint2 = midint
gsem(midint1 <- defz mcaplnk_0 mcaplnk_1 mcaplnk_2 mcaplnk_3 mcaplnk_4 mcaplnk_5 mcaplnk_6 mcaplnk_7
     c.defz#c.mcaplnk_0 c.defz#c.mcaplnk_1 c.defz#c.mcaplnk_2  /*
*/ c.defz#c.mcaplnk_3 c.defz#c.mcaplnk_4 c.defz#c.mcaplnk_5 c.defz#c.mcaplnk_6 c.defz#c.mcaplnk_7
     M1[ddyadid], logit) (midint2 <- defz mcaplnk_0 mcaplnk_1 mcaplnk_2 mcaplnk_3 /*
*/ mcaplnk_4 mcaplnk_5 mcaplnk_6 mcaplnk_7 c.defz#c.mcaplnk_0 c.defz#c.mcaplnk_1 c.defz#c.mcaplnk_2
     c.defz#c.mcaplnk_3 c.defz#c.mcaplnk_4 c.defz#c.mcaplnk_5 /*
*/ c.defz#c.mcaplnk_6 c.defz#c.mcaplnk_7 pceyrsk_1 pceyrsk_2 pceyrsk_3 pceyrsk_4 M2[ddyadid],logit),
     cov(M1[ddyadid]*M2[ddyadid]@0) vce(robust)

*** REWB GSEM
clonevar midint3 = midint
clonevar midint4 = midint
gsem(midint3 <- W_defz B_defz mcaplnk_0 mcaplnk_1 mcaplnk_2 mcaplnk_3 mcaplnk_4 mcaplnk_5 mcaplnk_6
     mcaplnk_7 W_defXR_mcaplnk_0 W_defXR_mcaplnk_1 W_defXR_mcaplnk_2 /*
```

```
*/ W_defXR_mcaplnk_3 W_defXR_mcaplnk_4 W_defXR_mcaplnk_5 W_defXR_mcaplnk_6 W_defXR_mcaplnk_7
    B_defXR_mcaplnk_0 B_defXR_mcaplnk_1 B_defXR_mcaplnk_2 B_defXR_mcaplnk_3 /*
*/ B_defXR_mcaplnk_4 B_defXR_mcaplnk_5 B_defXR_mcaplnk_6 B_defXR_mcaplnk_7 M1[ddyadid], logit) (midint4
    <- W_defz B_defz mcaplnk_0 mcaplnk_1 mcaplnk_2 mcaplnk_3 /*
*/ mcaplnk_4 mcaplnk_5 mcaplnk_6 mcaplnk_7 W_defXR_mcaplnk_0 W_defXR_mcaplnk_1 W_defXR_mcaplnk_2
    W_defXR_mcaplnk_3 W_defXR_mcaplnk_4 W_defXR_mcaplnk_5 W_defXR_mcaplnk_6 /*
*/ W_defXR_mcaplnk_7 B_defXR_mcaplnk_0 B_defXR_mcaplnk_1 B_defXR_mcaplnk_2 B_defXR_mcaplnk_3
    B_defXR_mcaplnk_4 B_defXR_mcaplnk_5 B_defXR_mcaplnk_6 B_defXR_mcaplnk_7 /*
*/ pceyrsk_1 pceyrsk_2 pceyrsk_3 pceyrsk_4 M2[ddyadid], logit), cov(M1[ddyadid]*M2[ddyadid]@0)
    vce(robust) nocapslatent
```

Skills 1.4.2
See above.

Figure 1.4.3: *Plot of the within effect of status deficit across time with mean and in-sample binned estimates*



Note. Program to generate in-sample binned estimates is highly computationally intensive. For now, I present estimates derived from population samples to speed up computation time.

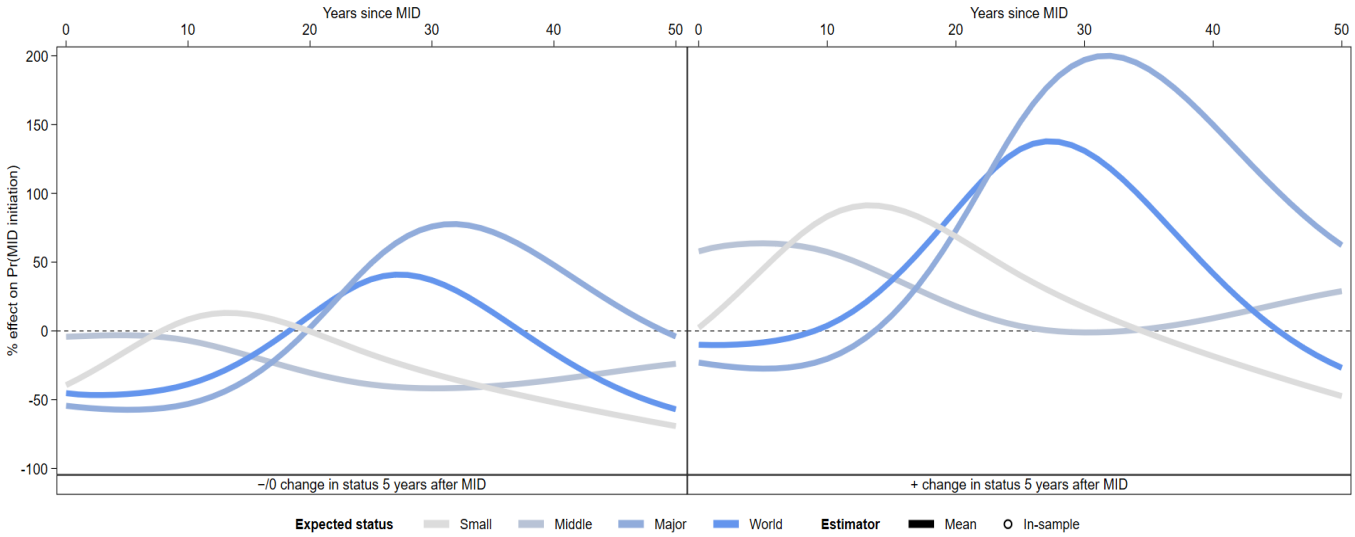
Stata code 1.4.3: *"By-hand" computation of the substantive effect of within-country status deficit and cross-level interaction terms.*

```
*** Loop for discrete change in status deficit
    local f 0
    qui forvalues p = 0/1 {
        local ++ f
        if `f' != 1 {
        *| Re-generate within/between status deficit components at + 1. S.D.
            *| for 1st ob. only so as to preserve within-country variation
        *| NOTE. 1st ob only bc. if all obs = 1, then within effect (cluster deviation) = 0
            qui replace defz = `p' if `obno' == 1
            *| Re-generate within/between components at +1 S.D. (for 1st ob.)
            by ddyadid: center defz , prefix(W2_) mean(B2_)
        *| Update within/between components at +1 S.D. values (for 1st ob.)
        replace W_defz = W2_defz[1] // within = +1 S.D.
        replace B_defz = B2_defz[1] // between = mean (0)
         }
          *| Update within-deficitXbetween-CINC(ln) smoothed interaction
        forvalues k = 0/7 {
            *| For status deficit = 0, use base variable (W_defz where within effect = 0)
                if `f' == 1 {
                gen double W2_defXB_mcaplnk_`k' = W_defz*B_mcaplnk_`k'
            }
            *| For status deficit = 1, use re-generated variable (W2_defz where within effect for 1st ob.
                 = 1)
            else {
                gen double W2_defXB_mcaplnk_`k' = W2_defz*B_mcaplnk_`k'
```

```
    }
    *| Update within and between components of interaction
    *| NOTE. Standard multiplicativive interaction Xit*Zi(hat) is not valid!!!!!!!!!!!!!!
    *| Correct specification = XitZi(hat)it*XitZi(hat)i(hat)
    by ddyadid: center W2_defXB_mcaplnk_`k', prefix(W2_) mean(B2_)
    su W2_W2_defXB_mcaplnk_`k' if `obno' == 1, meanonly
    replace W_W_defXB_mcaplnk_`k' = r(mean)
    su B2_W2_defXB_mcaplnk_`k' if `obno' == 1, meanonly
    replace B_W_defXB_mcaplnk_`k' = r(mean)
    *| Update 3-way interaction (within-deficitXbetween-CINC(ln) smoothedXpeace years
    forvalues k2 = 1/4 {
        replace W_W_defXB_mcaplnk_`k'Xpyk_`k2' = W_W_defXB_mcaplnk_`k'*pceyrsk_`k2'
            replace B_W_defXB_mcaplnk_`k'Xpyk_`k2' = B_W_defXB_mcaplnk_`k'*pceyrsk_`k2'
        }
    }
    *| Generate predictions with partial derivatives
    if `f' == 1 {
    qui predictnl double prm = predict(pr), g(dm_)
}
```

Skills 1.4.3
Modelling/computing/illustrating the combined effect of material and time dependence via REWB cross-level three-way interaction terms.
Calculating/plotting of in-sample binned estimates to show variation around the mean effect.

Figure 1.4.4: *Plot of the within effect of status deficit across time by past experience of war.*



Note. Program to generate in-sample binned estimates is highly computationally intensive. For now, I present mean estimates only.

Stata code 1.4.4: *Program for multi-core computation of in-sample estimates.*

```
***Set Stata to run across multiple cores (to drastically speed up computation time)
parallel initialize 6, force s("C:\Program Files\Stata17\StataBE-64.exe")
*** Program for multi-core computation of in-sample predictions
capture program drop bytimeg2plus
program define bytimeg2plus
    syntax varlist
    *| Sort by dyad sample size (to ensure computation of within effect on full 52 year sample)
    tempname dyobs
    bysort ddyadid: gen `dyobs' = _N
    gsort - `dyobs' +ddyadid + year
    *| Tag 1st ob. (always full sample dyad)
    gen obno1 = _n
    *| Set status deficit at +1 S.D. for 1st ob. only
    qui replace defz = 1 if obno1 == 1
    *| Macro for 1st ob. position when re-sorted by dyad-year
    sort ddyadid year
    tempname neworder
    gen `neworder' = _n
    su `neworder' if obno1 == 1, meanonly
    local ob1pos = r(mean)
    *| Re-generate within status deficit components at + 1 S.D. (for 1st ob)
    by ddyadid: center defz , prefix(W2_) mean(B2_)
```

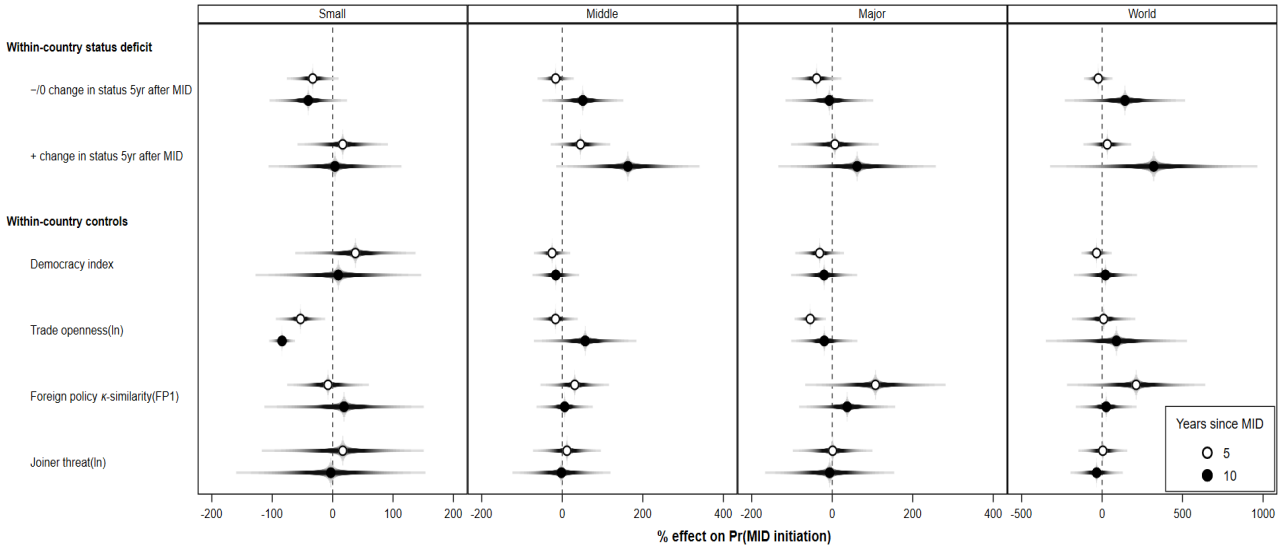
```
        *| Temp file to post estimation results
        estimates use prs
        tempname plus
        postfile `plus' ddyadid year _prp pceyrs using rpe_bytimeg2_plus_`varlist', replace
        *| Loop over each ob in sample
        forvalues i = 1/_N{
            *| Set restoration point
            preserve

            ********** INSET ESTIMATION SCRIPT FROM Figure 1.4.3 **********

            *| Generate predictions
            predict double pr`i', pr
            *| Post prediction and ob. identifiers to tempfile
            local ddyadid = ddyadid[`i']
            local year = year[`i']
            local _prp = pr`i'[`ob1pos']
            local pceyrs = pceyrs[`i']
            post `plus' (`ddyadid') (`year') (`_prp') (`pceyrs')
            restore
        }
        *| Save tempfile
        postclose `plus'
    end
```

Skills 1.4.4
Inference/modelling of time-dependence across two dimensions.
+ See above.

Figure 1.4.5: *Coefficient plot of the % effect of within-country status deficit versus within-country controls.*



Note. Coefficient plot of the discrete percentage effect of status deficit/control variables on the probability of initiating an MID across levels of expected status. Estimates are within-country effects drawn from a REWB model with status deficit, CINC(ln) smoothed, peace years smoothed and four additional control variables. A three-way, cross-level interaction is also included between *a*) the within-effect of status deficit/each control, *b*) the between effect of CINC(ln) smoothed and *c*) the random effect of peace years smoothed. Estimates represent the effect of a 1 unit increase, switching a variable's within component - and its corresponding interaction terms - from 0 to 1, while holding it's between component - and the within and between components of covariates - at the subpopulation (expected status) mean. Estimates are computed at two levels, fixing the dummy variable - status gains - at *a*) 0 and *b*) 1. The estimated effect of status deficit is plotted at both levels, with an interaction term between status deficit and status gains. Plotted control estimates represent the mean effect across levels of status gains. Estimates are also computed at two seperate temporal intervals. White and black circles represent the estimated effect for a state that last initiated an MID 5 and 10 years ago respectively. Confidence intervals are plotted across a range of confidence levels in gradating shades of color from dark (low) to light (high).

Stata code 1.4.5: *"By-hand" creation of Jacobian matrix for the calculation of confidence intervals via variance-covariance matrix.*

```
*** Generate Jacobian matrix (partial derivatives of logit predictions)
    *| Enables creation of VCV matrix for calculation of RR confidence intervals
    local rowno = `colno'    // Number of matrix cols. = no. predictions
    mat J = J(`rowno',`paramno',.)    // Create matrix
    local cnames_j: colnames e(V)   // Macro for matrix col. names
    mat colnames J = `cnames_j'   // Name matrix cols.
    local levs ng g // Macro for prediction type
    local k 0   // Increment
    foreach l in `levs' {    // Loop over no-gains/gains predictions
        local ++ k // Start increment
        forvalues i = 1/`paramno' {    // Loop over model paramters
            capture confirm var dm`l'_`i'    // Mean prediction
            if c(rc) == 111 {
                continue    // Skip the following if deriv. is missing
```

```
            }
            else {   // If deriv. is non-missing do the following
                if `k' == 1 {   // Where status gains = 0
                    mat J[1,`i'] = dm`l'_`i'[1]   // Deriv. for t=5
                    mat J[7,`i'] = dm`l'_`i'[2]   // Deriv. for t=10
                }
                else if `k' == 2 {   // Where status gains = 1
                    mat J[13,`i'] = dm`l'_`i'[1]   // Deriv. for t=5
                    mat J[19,`i'] = dm`l'_`i'[2]   // Deriv. for t=10
                }
            }
        }
        forvalues j = 1/5 {   // Loop over main effects (status deficit / controls)
            forvalues i = 1/`paramno' {   // Loop over model paramters
                capture confirm var dp`l'_`j'_`i'   // +1 S.D. prediction
                if c(rc) == 111 {
                    continue   // Skip the following if deriv. is missing
                }
                else {   // Where deriv. is non-missing do the following
                    if `k' == 1 {   // Where gains = 0
                        local jp1 = `j' + 1   // Macro for matrix row no.
                        mat J[`jp1',`i'] = dp`l'_`j'_`i'[1]   // Deriv. for t=5
                        local jp2 = `j' + 7 // Macro for matrix row no.
                        mat J[`jp2',`i'] = dp`l'_`j'_`i'[2]   // Deriv. for t=10
                    }
                    else if `k' == 2 {   // Where gains = 1
                        local jp3 = `j' + 13   // Macro for matrix row no.
                        mat J[`jp3',`i'] = dp`l'_`j'_`i'[1]   // Deriv. for t=5
                        local jp4 = `j' + 19 // Macro for matrix row no.
                        mat J[`jp4',`i'] = dp`l'_`j'_`i'[2]   // Deriv. for t=10
                    }
                }
            }
        }
    }

    forvalues i = 1/`rowno' {   // Loop over predictions
        forvalues j = 1/`paramno' {   // Loop over parameters
            if missing(J[`i', `j']) {   // Where deriv. is missing...
                matrix J[`i', `j'] = 0   // ... Deriv. = 0
            }
        }
    }

    *** Generate variance-covariance matrix
    mat V = J*e(V)*J'
```

Skills 1.4.5
Computation of Cohens kappa cluster-similarity algorithm in the creation of foreign policy similarity variable. Modelling/computing the effects of 350+ parameters (approx. 70 per main effect)

2 Data Analytics at Costcutter Canterbury

An on-going business initiative utilising SQL, Power BI and R

2.1 Project summary

Aims
To challenge the existing business model via the utilisation of data management tools and descriptive/prescriptive data-driven insights.

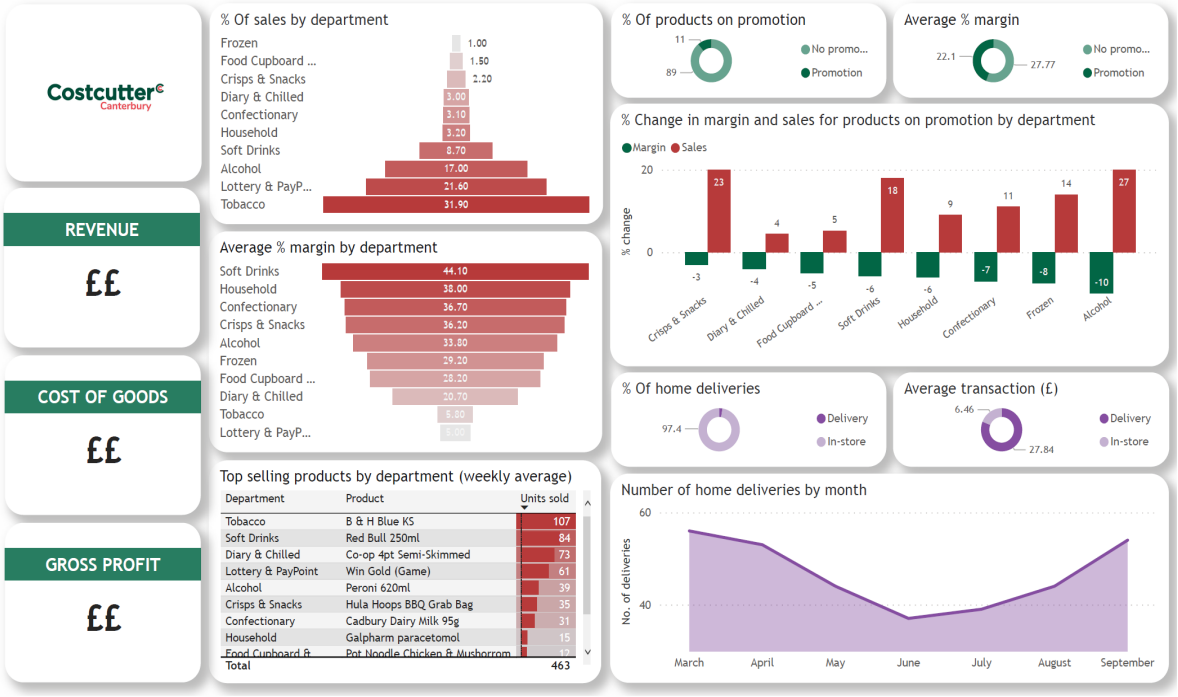
Methods
Demonstrating the added-value of data collection/analysis to the convenience store operation in the following two ways:
1) I produced functional MS excel spreadsheets and descriptive Power BI reports, giving management greater control and understanding of inventory, revenue and operations.
2) I collected data on in-store and home-delivery custom via survey designs and SQL database extraction, which I used to test various business hypotheses via statistical analysis.

Outcomes
I successfully initiated and led the digitisation of the convenience store operation as well as delivering business insights that informed, for example, targeted social media advertisement and the continued operation on an in-store ATM.

Skills
Identifying an opportunity to advance the business and apply my statistical knowledge to a business setting.
Demonstrating the value of data-driven insights via highly accessible data visualisations and clear verbal explanation of statistical concepts.
Employing a variety of standard and advanced techniques in MS Excel, SQL, Power BI and Stata.

2.2 Data management and descriptive analysis

Figure 2.2.1: *Power BI report for Costcutter Canterbury*



SQL code 2.2.1: *Query to generate data on the average % margin by department.*

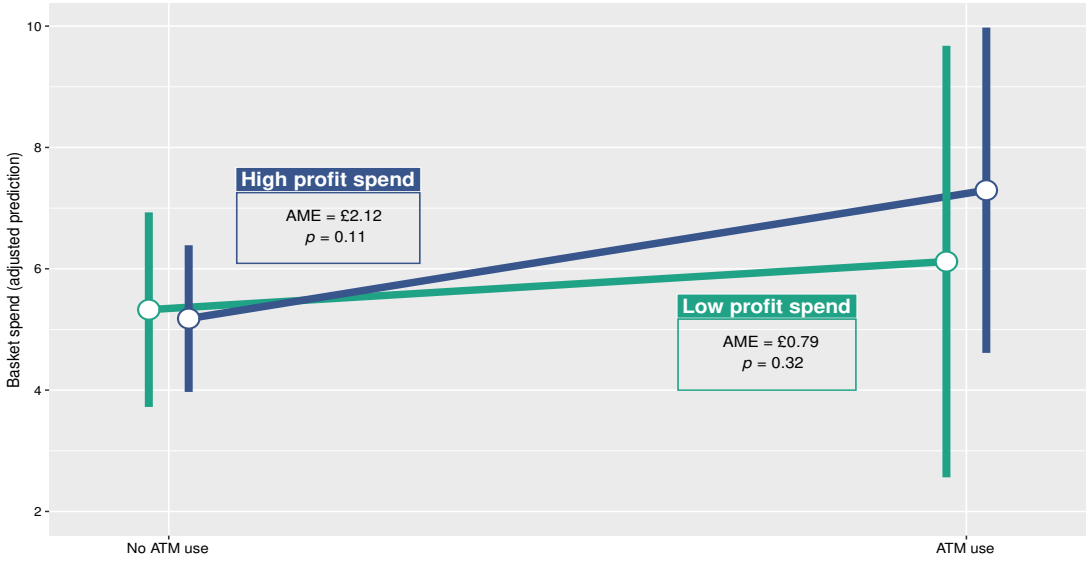
```
/* AVERAGE MARGIN BY DEPARTMENT */
/*****************************************************************/
/* Get updated department names */
DROP VIEW IF EXISTS ProductsNewDep
GO
CREATE VIEW ProductsNewDep AS
    SELECT
        DepDictionary.DepartmentNew,
        DepDictionary.Department,
        tblProducts.Stocked,
        tblProducts.SupplierCost,
        tblProducts.PackQuantity,
        tblProducts.CurrentSell,
        tblProducts.WeeklySales,
        tblProducts.Description
    FROM
        DepDictionary
    INNER JOIN
        tblProducts ON tblProducts.Department=DepDictionary.Department;
GO
```

```
/* Filtered table */
DROP TABLE IF EXISTS ProductsNewDepPrelim
SELECT
    Department,
    CAST(DepartmentNew AS INT)[DepartmentNew],
    SupplierCost,
    (PackQuantity*CurrentSell) AS Revenue,
    Description,
    WeeklySales
    INTO
         ProductsNewDepPrelim
    FROM
         ProductsNewDep
    WHERE
         Stocked = 1 AND DepartmentNew IS NOT NULL AND SupplierCost > 0
         AND NOT(Description = 'LIPTON ICE TEA PEACH PM1') AND NOT(Description = 'DIET COKE PM1.75')
         AND NOT(Description = 'MOUNTAIN DEW REGULAR PM1.15') AND NOT(Description = 'PEPSI REGULAR')
GO
/* Results table  */
DROP TABLE IF EXISTS DepAvMar
GO
SELECT
    DepartmentNew,
    ROUND(AVG(((Revenue-SupplierCost)/NULLIF(Revenue,0))*100),1) AS AvMargin
INTO
    DepAvMar
FROM
    ProductsNewDepPrelim
GROUP BY
    DepartmentNew
ORDER BY
    DepartmentNew
GO
```

Skills 2.2.1
Employing advanced power bi techniques such as conditional formatting and DAX functions.
Generating custom datasets via functional SQl queries

2.3 Delivering actionable business insights via statistical analysis

Figure 2.3.1: *Plot of the average marginal effect (AME) of ATM use on basket spend.*



Skills 2.3.1
Using in-text signposting instead of legend to maximise readability.

Figure 2.3.2: *Plot of the discrete effect of age on basket spend.*

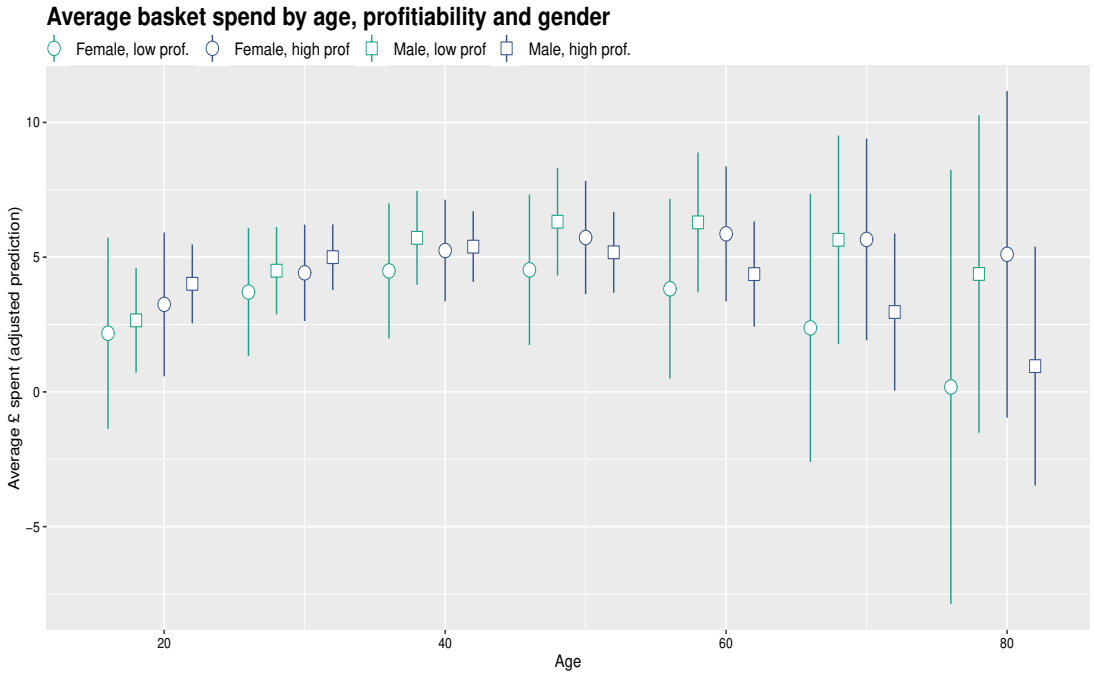
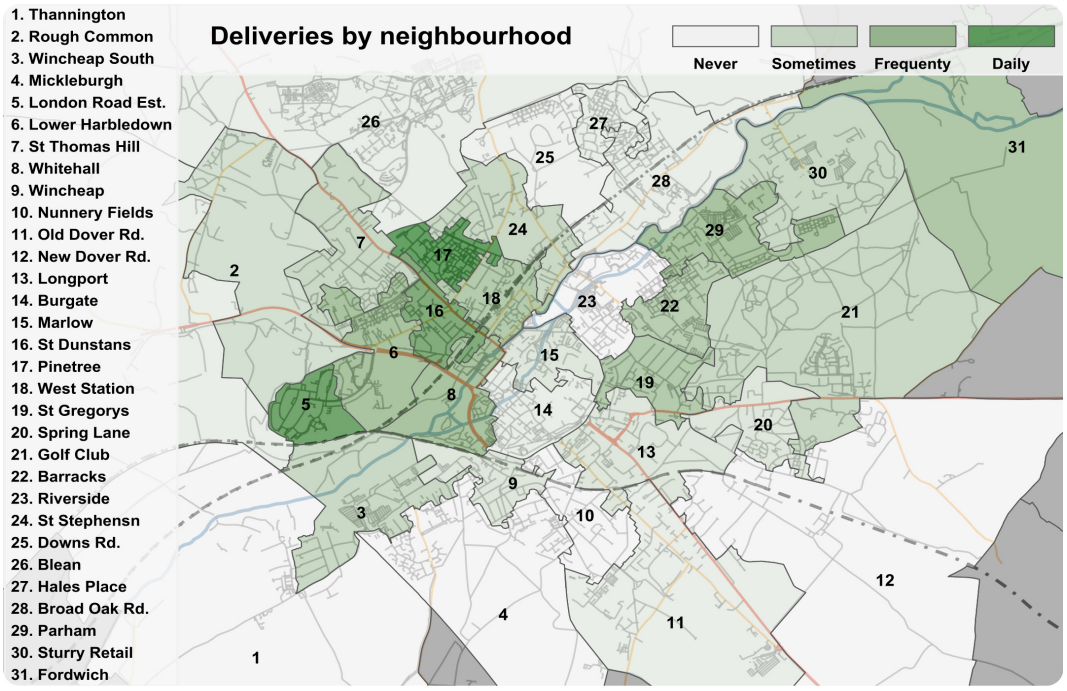


Figure 2.3.3: *Geospatial heatmap of deliveries by neighbourhood.*



R code 2.3.3: *Geospatial plot using Open Street Map data and LSOA polygons.*

```
### Plot
delmap <- ggplot() +                    #ggplot environment
  geom_sf(data = river$osm_lines,       # add river
          inherit.aes = FALSE,
          color = "steelblue",
          size = .8,
          alpha = .8) +
  geom_sf(data = railway$osm_lines,     # add railway
          inherit.aes = FALSE,
          color = "black",
          size = .7,
          linetype="dotdash",
```

```
        alpha = .8) +
geom_sf(data = med_streets$osm_lines,  #add streets
        inherit.aes = FALSE,
        color = "#feb24c",
        size = .5,
        alpha = .8) +
geom_sf(data = small_streets$osm_lines, # add roads
        inherit.aes = FALSE,
        color = "#666666",
        size = .4,
        alpha = .65) +
geom_sf(data = big_streets$osm_lines,  # add main roads
        inherit.aes = FALSE,
        color = "#f03b20",
        size = .6,
        alpha = .8) +
geom_sf(data = lsoa_shape_delno,       # add shape polygons
        aes(fill = delno),
        alpha = 0.6,
        size = .25,
        inherit.aes = FALSE)+
scale_fill_gradient(low='#EEEEEE',     # gradient box as custom legend
                    high='darkgreen',
                    breaks=c(0, 4, 8.25, 12),labels=c("Never","Sometimes", "Frequenty","Daily")) +
geom_sf_text(data = lsoa_shape,        # LSOA no. labels
             aes(label = lsoalabs),
             size = 2.95, fontface = 2) +
annotate("text", x= 1.063, y = 51.296,    # confine wayward label
label = "26", size = 2.95, fontface = 2) +
coord_sf(xlim = c(1.035, 1.1225),         # axis range
         ylim = c(51.26, 51.3035),
         default_crs = sf::st_crs(4326),
         expand = FALSE, clip = "off") +
annotate("rect", xmin = 1, xmax = 1.04425, # border box
         ymin = 51.26, ymax=51.75, fill = "grey98",
         alpha = 0.8, color="grey30") +
annotate("rect", xmin = 1.04425, xmax = 2, # border box
         ymin = 51.299, ymax=52, fill = "grey98",
         alpha = 0.8, color="grey30") +
annotate("text", x= 1.0281, y = 51.281782, # LSOA names
         label = "1. Thannington \n2. Rough Common \n3. Wincheap South\n4. Mickleburgh\n5. London
              Road Est.\n6. Lower Harbledown\n7. St Thomas Hill\n8. Whitehall\n9. Wincheap\n10.
              Nunnery Fields\n11. Old Dover Rd.\n12. New Dover Rd.\n13. Longport\n14. Burgate\n15.
              Marlow\n16. St Dunstans\n17. Pinetree\n18. West Station\n19. St Gregorys\n20. Spring
              Lane\n21. Golf Club\n22. Barracks\n23. Riverside\n24. St Stephensn\n25. Downs Rd.\n26.
              Blean\n27. Hales Place\n28. Broad Oak Rd.\n29. Parham\n30. Sturry Retail\n31.
              Fordwich",
         size = 2.75, fontface = 2,  hjust = 0) +
annotate("text", x= 1.065, y = 51.3015,    # legend title
         label = "Deliveries by neighbourhood", size = 4.5, fontface = 2) +
guides(fill = guide_legend(                # legend alignment
  direction = "horizontal",
  label.position = "bottom"))+
theme_void() +                              # theme
theme(legend.position=c(.86, .935)) +       # legend position
theme(legend.title = element_blank()) +     # legend title off
theme(legend.key.height = unit(.4, 'cm')) + # symbol height
theme(legend.key.width = unit(1.5, 'cm'))+  # symbol width
theme(legend.text=element_text(size=7.75, face = 2)) + # text size
theme(text = element_text(family = "Roboto")) #font
```

Skills 2.3.3
Utilising advanced geospatial data visualisation methods.
Innovative use gradient text boxes to circumvent restrictive ggplot legend options.

3 A Method to Enable Time Series Analysis of Multiple Deprivation Indices with an Illustrative Analysis of Mortality Rates

A research paper based on analysis in R

3.1 Project summary

Aims
To devise a method to enable time series analysis of Multiple Deprivation Indices (IMD) - the standard measure of localised deprivation within the UK.

Methods
Enabling cross-year comparability between year-standardized IMD scores via coefficient/residual scaling. The analysis is structured as follows:
1) I show that the weighting - and underlying indicators - of four out of seven domains of IMD are constant across time, enabling the rescaling of IMD score for year *t* by the coefficients and residuals from an OLS regression of IMD score (on raw sub-domain indicators) for *t*+5
2) I will illustrate the potential of this method via time series analysis of the effect of within-neighbourhood deprivation on mortality rate using Random Effects Within Between models.

Skills
Utilising and devising advanced statistical method.
Utilising advanced table creation and plotting techniques.

3.2 Creating a time-adjusted IMD

Table 3.2.1: *By-year comparison of the underlying indicators of four domains of IMD*

	IMD 2015	IMD 2019
Education	13.5% [a]	13.5%
Key Stage 2 attainment [†]	✓	✓
Key Stage 4 attainment [†]	✓	✓
Secondary school absence [†]	✓	✓
Staying on in education post 16	✓	✓
Entry to higher education	✓	✓
Adults with no or low qualifications [††]	✓	✓
English langauge proficiency [††]	✓	✓
Health and Disability	13.5%	13.5%
Years of potential life lost	✓	✓
Comparative illness and disability ratio [b]	✓	✓
Acute morbidity	✓	✓
Mood and anxiety disorders [c]	✓	✓
Barriers to Housing and Services	9.3%	9.3%
Road distance to a:	✓	✓
• post office	✓	✓
• primary school	✓	✓
• general store	✓	✓
• GP surgery	✓	✓
Household overcrowding	✓	✓
Homelessness	✓	✓
Housing affordability	✓	✓
Living Environment	9.3%	9.3%
Houses without central heating	✓	✓
Housing in poor condition	✓	✓
Air quality	✓	✓
Road traffic	✓	✓

[a] % represents the weight attributed to each domain within the IMD.
[b] The introduction of Universal Credit (UC) and Personal Independence Payments (PIP) post IMD 2015 does not appear to have meaningfully altered the conditionality of health-related benefit claims for the 2019 IMD given that *1*) only a very small number of those eligible for health-related benefits were receiving UC for the data period (March 2016) on which the indicator is based and *2*) eligibility conditions for PIP and the Disability Living Allowance benefit it replaced are consistent on core dimensions i.e. non-means tested, non-taxable and payable to people who are in or out of work.
[c] The benefit-based component of this indicator was dropped for IMD 2019 due to concerns around data quality. No changes were made to its other three components.
[†] Data for this indicator is not published. [††] Data for this indicator is published as a combined indicator.

LaTeX***code 3.2.1:*** *Partial construction of table plus tick/line objects*

```
% Define table objects
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
\def\checkmark{\tikz\fill[scale=0.4](0,.35) -- (.25,0) -- (1,.7) -- (.25,.15) -- cycle;}  % Checkmark symbol
\newcommand{\greyrule}{\arrayrulecolor{black!30}\midrule\arrayrulecolor{black}} % Grey mid-rule lines
% Table environment
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
\begin{tab}[!htp] % Set custom floating environment
    \scriptsize % Fontsize
    \noindent % Does what it says on the tin
```

```
\centering % And again
\caption{\raggedright{By-year comparison of the underlying indicators of four domains of IMD}} % Caption
\begin{threeparttable} % 3-part table for easy manipulation of table notes
    \begin{tabular}{ p{7cm\textwidth} p{2.5cm\textwidth} p{2.5cm\textwidth}  } % Fix column widths
        \toprule % Ruling
        &\textbf{IMD 2015} &\textbf{IMD 2019} \\ % Sub-headings
        \midrule % More ruling
        % Domains
        %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
        \textbf{\emph{Education}}     &13.5\% \tnote{a}  &13.5\%  \\ % Domain heading
        \hspace{14pt} Key Stage 2 attainment \tnote{\textdagger} &\checkmark &\checkmark \\  % Checkmarks
            and so forth...
        \hspace{14pt} Key Stage 4 attainment \tnote{\textdagger}  &\checkmark &\checkmark \\ %...
        \hspace{14pt} Secondary school absence \tnote{\textdagger}  &\checkmark &\checkmark \\ %...
        \hspace{14pt} Staying on in education post 16 &\checkmark &\checkmark \\ %...
        \hspace{14pt} Entry to higher education &\checkmark &\checkmark \\%...
        \hspace{14pt} Adults with no or low qualifications \tnote{\textdagger\textdagger} &\checkmark
            &\checkmark \\ %...
```

Table 3.2.2: *Table of coefficients from linear and multivariable fractional polynomial (MFP) least squares regressions of IMD score on IMD sub-domain scores by year*

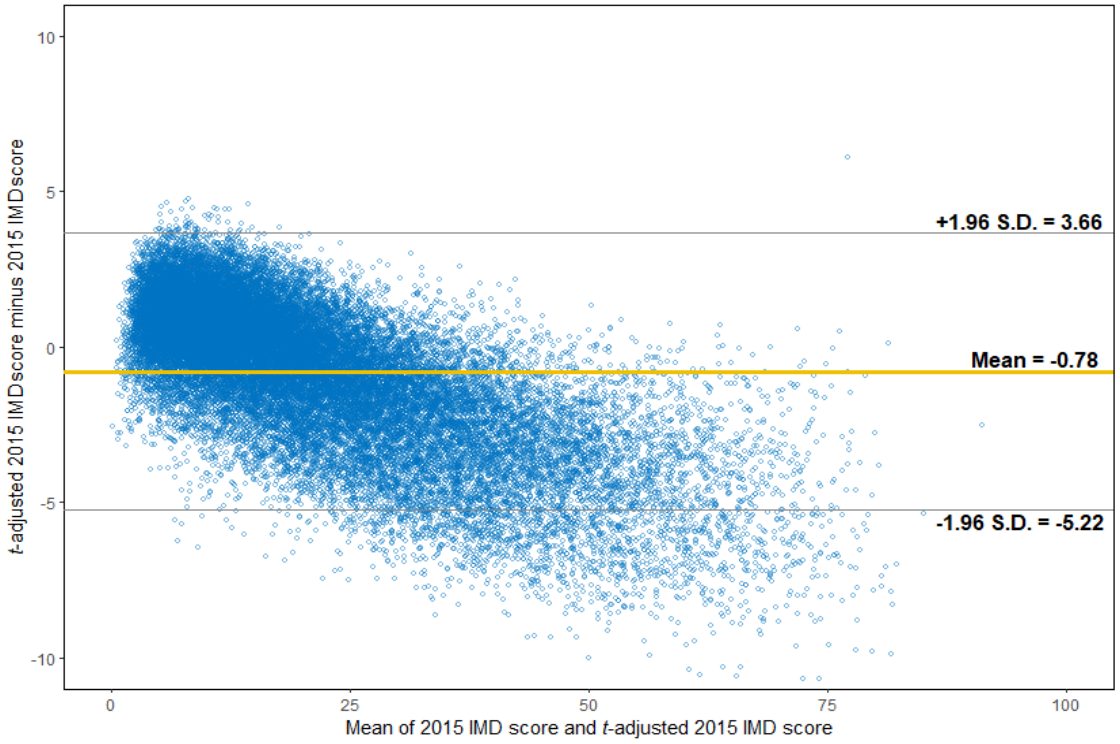
Sub-domain indicators	IMD 2015		IMD 2019	
	Linear	MFP	Linear	MFP
Education				
Post_16	0.33***	-	0.03***	-
FP1	-	0.09***	-	0.07***
FP2	-	-0.07***	-	-0.05***
Higher	-0.03***	-	-0.02***	-
FP1	-	-0.14***	-	-0.07***
FP2	-	0.17***	-	0.08***
Adult	0.38***	-	0.28***	-
FP1	-	-0.02***	-	-0.01***
FP2	-	0.36***	-	0.31***
Health				
Life_lost	0.19***	-	2.31e-4	-
FP1	-	-0.25***	-	-0.14***
FP2	-	-0.24***	-	-0.14***
Ill_dis_ratio	0.44***	-	0.65***	-
FP1	-	0.84***	-	1.38***
FP2	-	-0.51***	-	-0.91***
Acu_morbid	0.07***	-	0.05***	-
FP1	-	-0.35***	-	-0.2***
FP2	-	-0.31***	-	-0.15***
Mood_anxi	0.06***	-	4.1e3	-
FP1	-	0.02***	-	0.03***
FP2	-	0.09***	-	0.03***
Barriers				
PO_dist	0.03***	-	0.03***	-
FP1	-	0.03***	-	0.02***
Primary_dist	0.03***	-	0.03***	-
FP1	-	0.09***	-	0.07***
FP2	-	-0.06***	-	-0.05***
Shop_dist	0.03***	-	0.02***	-
FP1	-	0.08***	-	0.06***
FP2	-	0.11***	-	-0.08***
GP_dist	0.02***	-	0.02***	-
FP1	-	0.03***	-	0.03***
FP2	-	-7.4e3***	-	-9.1e3***
Overcrowding	0.15***	-	0.08***	-
FP1	-	-0.71***	-	-0.48***
FP2	-	-0.54***	-	-0.33***
Homelessness	0.05***	-	0.08***	-
FP1	-	0.07***	-	0.05***
Environment				
Poor_cond	0.04***	-	0.05***	-
FP1	-	0.05***	-	0.06***
No_heating	0.07***	-	0.08***	-
FP1	-	0.17***	-	0.21***
FP2	-	-0.11***	-	-0.14***
Raod_accid	3.2e3	-	0.01***	-
FP1	-	-0.04**	-	0.11***
FP2	-	0.0.05**	-	-0.08***
Air_qual	0.08***	-	0.04***	-
FP1	-	0.2***	-	0.07***
FP2	-	-0.26***	-	-0.08***
R^2	0.92	0.95	0.93	0.96
AIC	1906	1753	1829	1700

R code 3.2.2: *Linear and MFP regressions with FP transformations*

```
### Linear regression of IMD score on raw IMD sub-domain indicators
lm.2015 <- lm(IMD ~ educ_1 + educ_2 + educ_3 + health_1 + health_2 + health_3 + health_4 +
              barr_1 + barr_2 + barr_3 + barr_4 + barr_5 + barr_6 +
              livenv_1 + livenv_2 + livenv_3 + livenv_4, data = )
### MFP of IMD score on FP-transformed IMD sub-domain indicators (4 degrees of freedom i.e. max of FP2)
mfp2015 <- mfp(IMD ~ fp(educ_1) + fp(educ_2) + fp(educ_3) + fp(health_1) + fp(health_2) +
           fp(health_3) + fp(health_4) + fp(barr_1) + fp(barr_2) + fp(barr_3) + fp(barr_4) +
           fp(barr_5) + fp(barr_6) + fp(livenv_1) + fp(livenv_2) + fp(livenv_3) + fp(livenv_4),
           family = gaussian, data = , verbose = TRUE)
### FP transformations
mfp2015$formula
attach(IMD_raw_2015)
educ_1_FP1 <- (educ_1/0.1)^1
educ_1_FP2 <- log((educ_1/0.1))
educ_2_FP1 <- educ_2^3
educ_2_FP2 <- (educ_2^3)*log(educ_2)
educ_3_FP1 <- (educ_3/0.1)^-2
educ_3_FP2 <- (educ_3/0.1)^2
health_1_FP1 <- (health_1/100)^-1
health_1_FP2 <- (health_1^-1)*log((health_1/100))
health_2_FP1 <- (health_2/100)^3
health_2_FP2 <- ((health_2/100)^3)*log((health_2/100))
health_3_FP1 <- (health_3/100)^-1
health_3_FP2 <- ((health_3/100)^-1)*log((health_3/100))
health_4_FP1 <- log((health_4+2.9))
health_4_FP2 <- (health_4+2.9)^3
# Remaining transformations omitted #
detach(IMD_raw_2015)
```

Skills 3.2.1/3.2.2
Creating multi-column, three-part tables with custom objects.
Employing multivariable fractional polynomial procedure to maximise model fit.

Figure 3.2.3: *Mean difference plot of 2015 IMD score versus time-adjusted 2015 IMD score*



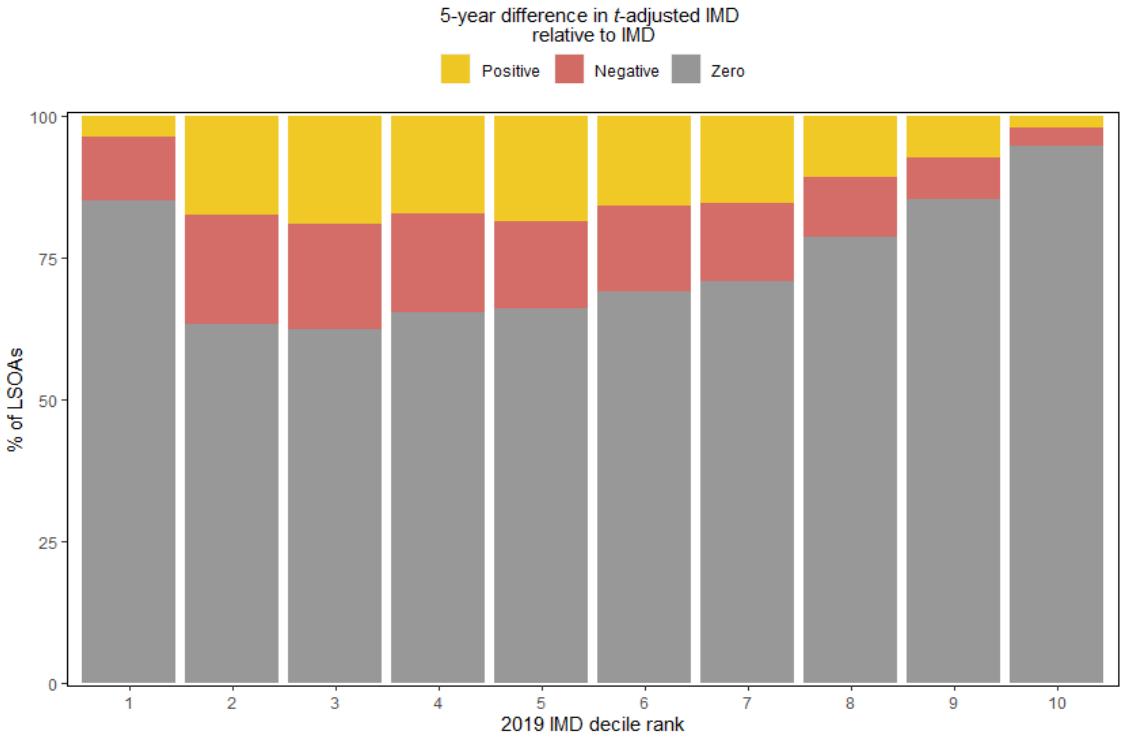
R code 2.3.3: *Mean difference plot*

```
### Get 2019 regression coefficients + 2015 residuals + residual SD for each year
mat_coef2019 <- lm.2019FP$coefficients
IMD_raw_2015$MFPres <- residuals(lm.2015FP)
resSD2015 <- sigma(lm.2015FP)
resSD2019 <- sigma(lm.2019FP)
### Create time-adjusted IMD
```

```
attach(IMD_raw_2015)
IMD_raw_2015$IMD_tadj <-
                        # sum of 2019 Intercept plus
                        (mat_coef2019[1] +
                        # 2015 indicator values scaled by 2019 coefficients plus
                        (mat_coef2019[2]*educ_1_FP1) + (mat_coef2019[3]*educ_1_FP2) +
                             (mat_coef2019[4]*educ_2_FP1) +
                        (mat_coef2019[5]*educ_2_FP2) + (mat_coef2019[6]*educ_3_FP1) +
                             (mat_coef2019[7]*educ_3_FP2) +
                        (mat_coef2019[8]*health_1_FP1) + (mat_coef2019[9]*health_1_FP2) +
                             (mat_coef2019[10]*health_2_FP1) +
                        (mat_coef2019[11]*health_2_FP2) + (mat_coef2019[12]*health_3_FP1) +
                             (mat_coef2019[13]*health_3_FP2) +
                        (mat_coef2019[14]*health_4_FP1) + (mat_coef2019[15]*health_4_FP2) +
                             (mat_coef2019[16]*barr_1_FP1) +
                        (mat_coef2019[17]*barr_2_FP1) + (mat_coef2019[18]*barr_2_FP2) +
                             (mat_coef2019[19]*barr_3_FP1) +
                        (mat_coef2019[20]*barr_3_FP2) + (mat_coef2019[21]*barr_4_FP1) +
                             (mat_coef2019[22]*barr_4_FP2) +
                        (mat_coef2019[23]*barr_5_FP1) + (mat_coef2019[24]*barr_5_FP2) +
                             (mat_coef2019[25]*barr_6_FP1) +
                        (mat_coef2019[26]*livenv_1_FP1) + (mat_coef2019[27]*livenv_2_FP1) +
                             (mat_coef2019[28]*livenv_2_FP2) +
                        (mat_coef2019[29]*livenv_3_FP1) + (mat_coef2019[30]*livenv_3_FP2) +
                             (mat_coef2019[31]*livenv_4_FP1) +
                        (mat_coef2019[32]*livenv_4_FP2) +
                        # 2015 standardized residual scaled by 2019 residual SD.
                        ((MFPres*resSD2019)/resSD2015))
```

Skills 3.2.3
Manual coding of regression scaling equation.

Figure 3.2.4: *Bar chart of the 5-year difference in IMD versus time-adjusted IMD*



R code 3.2.4: *Creation of time-adjusted IMD score*

```
### Plot
difindif<-ggplot(data=IMD_1519, aes(x=factor(IMD_decrk),
                                    y=DIF5ydif_perc, fill=direction, group=direction)) + # X/Y/by
          geom_bar(alpha = 0.85, stat="unique") + # Bars
          scale_fill_manual(labels = c("Positive ", "Negative", "Zero"), # Legend labels
                            values = c("1" = "#EDC000FF", # Yellow bars
                                       "2" = "#CD534CFF", # Red bars
                                       "3" = "#868686FF")) + # Grey bars
          scale_y_continuous(expand = c(.005,.005)) + # Reduce plot margin
          ylab('% of LSOAs') + # X-title
```

```
          xlab('2019 IMD decile rank')  + # Y-title
          guides(fill = guide_legend(title =
                                       '5-year difference in <i>t</i>-adjusted IMD<br>',
                                     title.position = "top")) + # Legend title in ggtext markdown
          labs(tag = 'relative to IMD') + # Second line of legend title as tag (to get around buggy
               positioning next to ggtext italic)
          coord_cartesian(xlim = c(1, 10), ylim = c(0, 100), clip = "off")+ # Allow tag to float
          theme(legend.position="top", legend.title.align=0.05, # Legend position
          panel.grid.major = element_blank(), # No grid lines
          panel.grid.minor = element_blank(), # As above
          panel.background = element_rect(
          colour = "black", size=.5, fill=NA), # Panel border
          legend.title=element_markdown(), # Declare markdown legend
          plot.tag.position = c(.53, .95),# Position tag as second line in legend title
          plot.tag=element_text(size=11)) # Same font size as legend title
difindif
```

Skills 3.2.4
In-text italics via ggmarkdown.
Use of text tag in legend to circumvent ggmarkdown buggy handling of multi-line text.