SCHOOL OF GEOGRAPHY UNIVERSITY OF LEEDS	UNIVERSITY OF LEEDS

COURSEWORK COVERSHEET

Student ID number	2	0	1	4	1	8	0	1	4
Module code	GEOG3195								
Module title	Geocomputation and Spatial Analysis								
Assignment title	GEOG3195 Assignment 2023/24 – GGP GAM								
Marker	Alexis Comber								
Declared word count	2493								

Your submission must comply with the University's definition of Academic Integrity as: "a commitment to good study practices and shared values which ensures that my work is a true expression of my own understanding and ideas, giving credit to others where their work contributes to mine". Double-check that your referencing and use of quotations is consistent with this commitment.

You must also ensure that your declared word count accurately reflects the number of words in your submission, excluding the overall title, bibliography/reference list, text/numbers in tables and figures (although table and figure captions are included in the word count).

An Analysis of COVID-19 Vaccination Rates Using GGP-GAM Modelling

1 Introduction

This paper creates and analyses a Generalized Additive Model with Gaussian Process splines (GGP-GAM), as proposed by Comber et al. (2023b), which investigates the effects of 4 key predictor variables (% over 65, % with level 4 qualifications, % in poor health, and % unemployed) on the COVID-19 vaccination rate in the UK.

A GGP-GAM was chosen due to its greater flexibility and therefore ability to handle localised patterns of spatial heterogeneity, which provide more accurate coefficient estimates in comparison to other SVCs such as an MGWR (Comber et al., 2023b). Notably, in this context, the GGP-GAM produces an impressive model fit with an adjusted R-squared of 0.917.

This paper outlines the steps required to construct a GGP-GAM and then provides an analysis of the coefficients produced by the model alongside visualisations created using mapping techniques. The discussion section of this paper provides an interpretation of the results and justifies the choice of a GGP-GAM before discussing the model tuning. Finally, the paper recognises the limitations of this research and outlines potential areas of further exploration.

2. Methods

2.1 Data Description

The vaccination data from the Coronavirus dashboard provides the target variable, the % of people vaccinated, over 6791 medium super output areas (MSOAs). The predictor variables were downloaded at MSOA level from NOMIS.

These data were joined to MSOA boundaries which were provided in a GeoPackage file to apply a geometry to the statistics from which a centroid was calculated to allow for spatial modelling.

2.2 Model Construction

The construction of a GGP-GAM required an intercept and an X and Y value to be added to the data frame. This was done using the `mutate` function in R. The longitudes and latitudes of MSOA centroids were assigned to X and Y, and the intercept was given a universal value of 1.

The GAM was then created using the `mgcv` package in R (Version 4.3.1).

```
gam.m <- gam(vacc ~ 0 +
             Intercept + s(X, Y, bs='gp', by=Intercept, k=k_choice) +
             o65 + s(X, Y, bs='gp', by=o65, k=k_choice) +
             l4qual + s(X, Y, bs='gp', by=l4qual, k=k_choice) +
             badhealth + s(X, Y, bs='gp', by=badhealth, k=k_choice) +
             unemp + s(X, Y, bs='gp', by=unemp, k=k_choice),
             data = df.gam)
```

Figure 1: R Code for model creation

In the code in Figure 1, the 0 is in the place of the intercept term, this can be seen as a dummy intercept since the predefined intercept is added later. The smooth term `s` is specified as the interaction between the coordinates (X, Y), conditioned by the predictor variables, using Gaussian Process as the basis. The smooth is controlled by specifying the number of knots (`k`) and this is how the model was tuned.

2.3 Model Tuning

`k` was pre-defined before running the code in Figure 1. After each model run a GAM check was run to assess the effects of increasing `k`. `k` was continually increased whilst the Hessian remained positive definite, and the smoothing parameter selection converged indicating stability. The goal of tuning the GAM using `k` was to increase the k-index to become close to 1 whilst maintaining some degree of computational efficiency. The model was run multiple times with a `k_choice` value between 40 and 180. In the case of this GGP-GAM, a `k_choice` of 180 began to produce diminishing returns in terms of k-index from a choice of 170, therefore 170 was selected as a fully tuned value for the smooth.

2.4 GGP-GAM SVC

A GGP-GAM SVC was constructed for each covariate by setting each variable to 1 and the others to 0. This was able to create smooth term estimates using the predict function as well as calculating the standard error.

2.5 Visualisation

Charts are created from the `gam.check` which aid in the visualisation of the model and its understanding. To visualise the results of the GGP-GAM the data was joined back to the geometry projection for mapping. The predicted values of the model were calculated and from these, a calculation of the residuals was created. The predicted vs actual values,

residuals, smooth term estimates, and standard errors were mapped using `ggplot2` in R and palettes from the `cols4all` library.

3. Results

3.1 GAM Summary

The model fit is validated by an adjusted R-squared value of 0.917 and 99.7% of the deviance is explained.

3.2 Coefficient Results

	Estimate	Std. Error	t value	Pr(>|t|)
s(X,Y):Intercept	0.444546	0.005021	88.540	<2e-16 ***
s(X,Y):o65	0.402094	3.041585	0.132	0.895
s(X,Y):l4qual	-0.158497	1.748935	-0.091	0.928
s(X,Y):badhealth	1.657690	10.967060	0.151	0.880
s(X,Y):unemp	2.598793	11.089195	0.234	0.815
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Figure 2: A table of the Parametric coefficients

The p-values in Figure 2 highlight the Intercept as being the only globally significant variable. The estimate of 0.444546 is recognised as the mean response from the model when all other variables = 0. In Figure 3 however, all smooth terms register a p-value of <0.05 highlighting their importance in representing non-linear relationships in the data, this can be viewed as being locally significant. The high F-value of l4qual (8.964) suggests that it adds strong explanatory power to the model.

	edf	Ref.df	F	p-value
s(X,Y):Intercept	93.60	96.39	7.032	<2e-16 ***
s(X,Y):o65	33.91	40.34	4.328	<2e-16 ***
s(X,Y):l4qual	32.02	40.22	8.964	<2e-16 ***
s(X,Y):badhealth	19.72	22.05	2.004	0.00411 **
s(X,Y):unemp	21.43	23.66	4.657	<2e-16 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Figure 3: A table of the approximate significance of smooth terms

3.3 General Visualisations

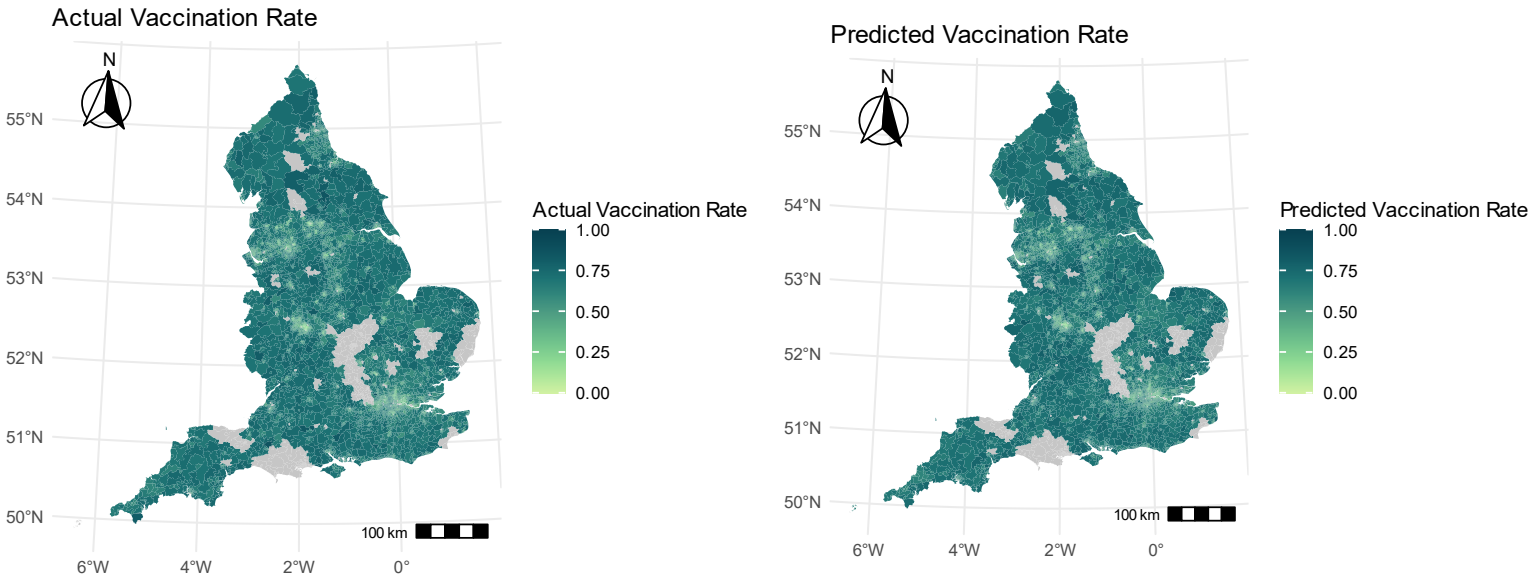

Figure 4: Maps showing the actual vaccination rate vs. predicted vaccination rate.

The maps in Figure 4 appear to recognise the same spatial patterns in predicted and actual vaccination rates which is expected due to the high R-squared value although a residuals map was created (figure 5) to aid in identifying any spatial concentrations of MSOAs for which the model did not predict as accurately.

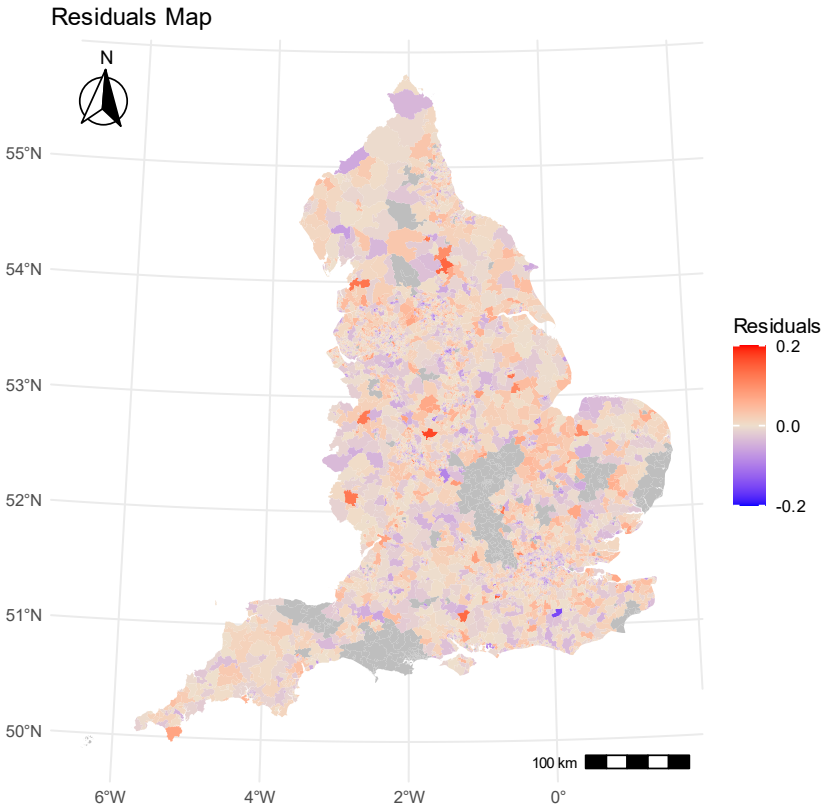

Figure 5: A map of the residuals from the model.

Figure 5 highlights the lack of spatial concentration in residuals suggesting that there is no problem with the underlying data or model creation. This can be further assessed by reviewing the standard error for each covariant.

3.4 Variable Visualisations

	Min	1st Quart	Median	3rd Quart	Max	IQR	Mean
b_Intercept	0.26	0.40	0.46	0.50	0.65	0.10	0.45
b_65	0.56	0.87	0.94	1.00	1.32	0.13	0.93
b_l4qual	0.00	0.14	0.19	0.29	0.47	0.15	0.22
b_badhealth	-1.29	-0.40	-0.10	0.09	0.49	0.49	-0.16
b_unemp	-2.92	-2.37	-2.05	-1.84	0.41	0.53	-2.06

Figure 6: Smooth term estimations table for all variables.

Since all variables are of local significance it is important to analyse the smooth term estimations. Figure 6 shows that an increase in the % of over 65's and % level 4 qualified leads to an increase in the vaccination rate. For these variables there is also a relatively small interquartile range between the estimations suggesting the spatial patterns are not as spatially concentrated as the bad health and unemployment variables. The bad health and unemployment variables also suggest the inverse with both being generally associated with a decrease in vaccination rate. This variation in coefficients can be viewed in Figures 7, 8, 9, 10 and 11.

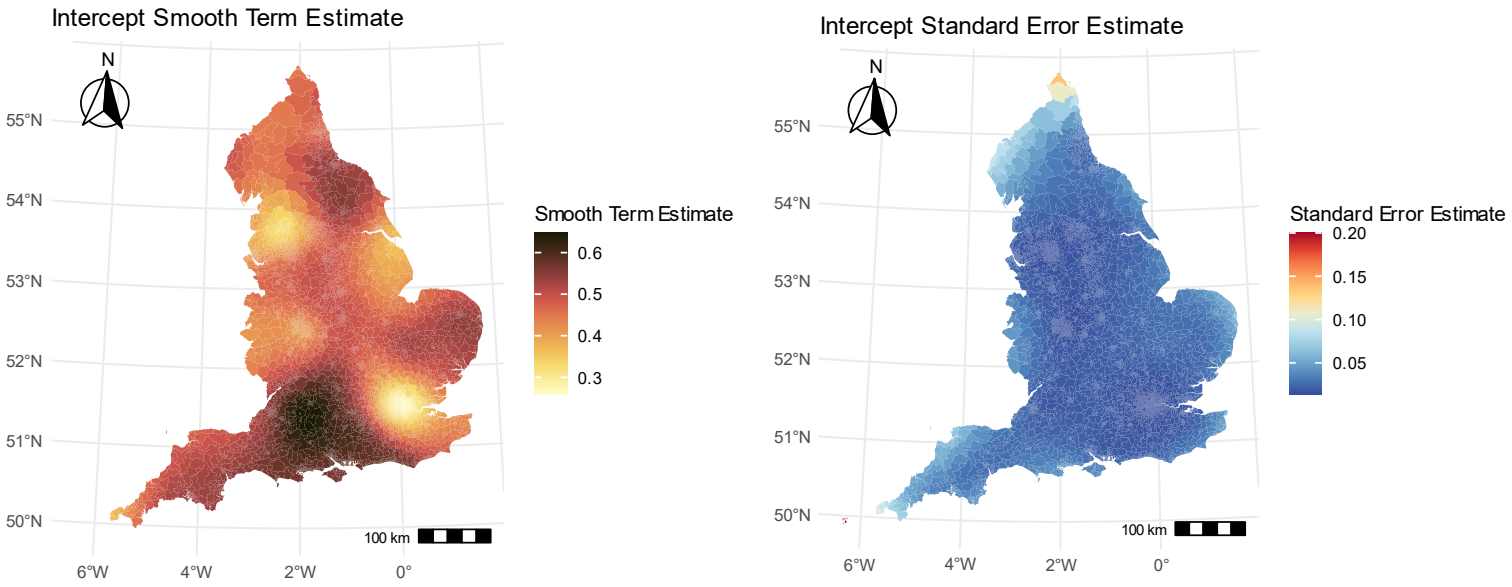

Figure 7: Maps showing the smooth term estimate and standard error for the Intercept.

The intercept has a lower smooth term estimate in London and the North-West as seen in Figure 7 suggesting a lower base rate of vaccination. The highest estimates fall within the South-West of the country.

There appears to be a generally low rate of standard error with only a small concentration of area close to the Scottish Border having a slightly higher rate.

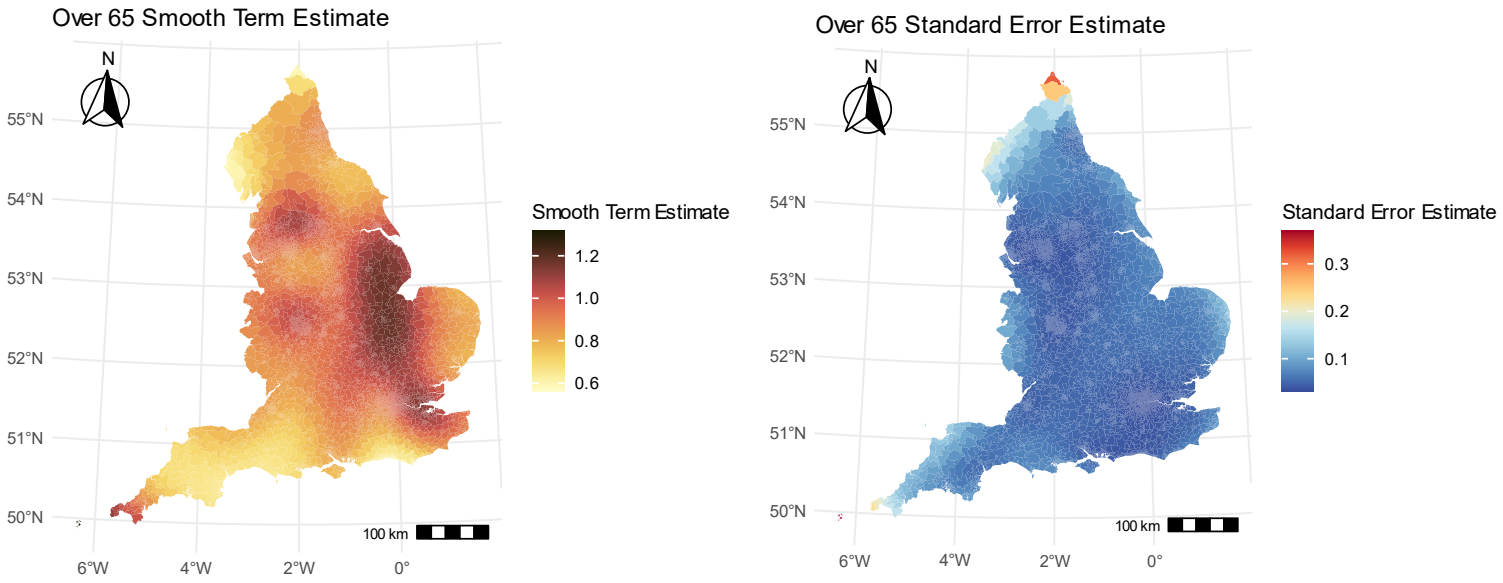

Figure 8: Maps showing the smooth term estimate and standard error for the over 65's Variable.

The % over 65 variable has the highest smooth term estimate in the east of England, therefore this is where it has the strongest impact on the model in terms of increasing the predicted vaccination rate in Figure 8. Again, standard error is low with a small concentration on the Scottish border.

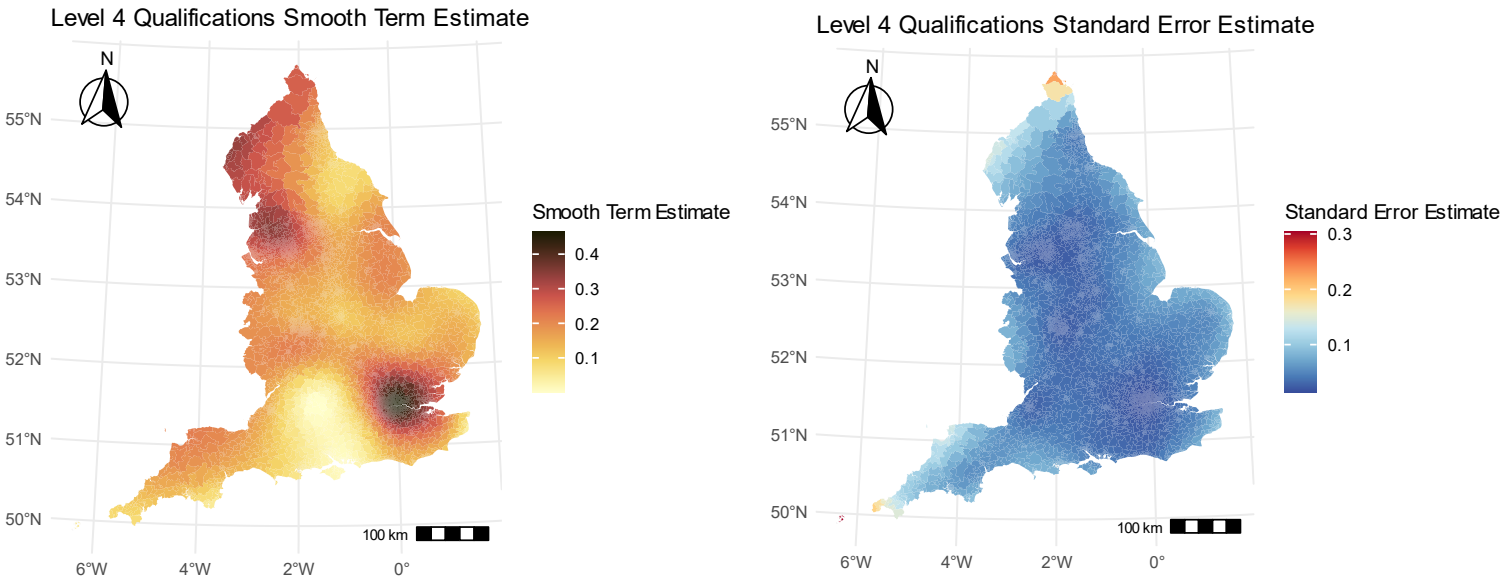

Figure 9: Maps showing the smooth term estimate and standard error for the Level 4 Qualifications Variable.

A strong spatial concentration can be recognised in London for the smooth terms estimate in Figure 9 with a higher rate of level 4 qualified individuals being a clear positive indicator of a higher vaccination rate. A low standard error rate is observed, with only small areas of higher rates visible along the Scottish border and in the South-East.

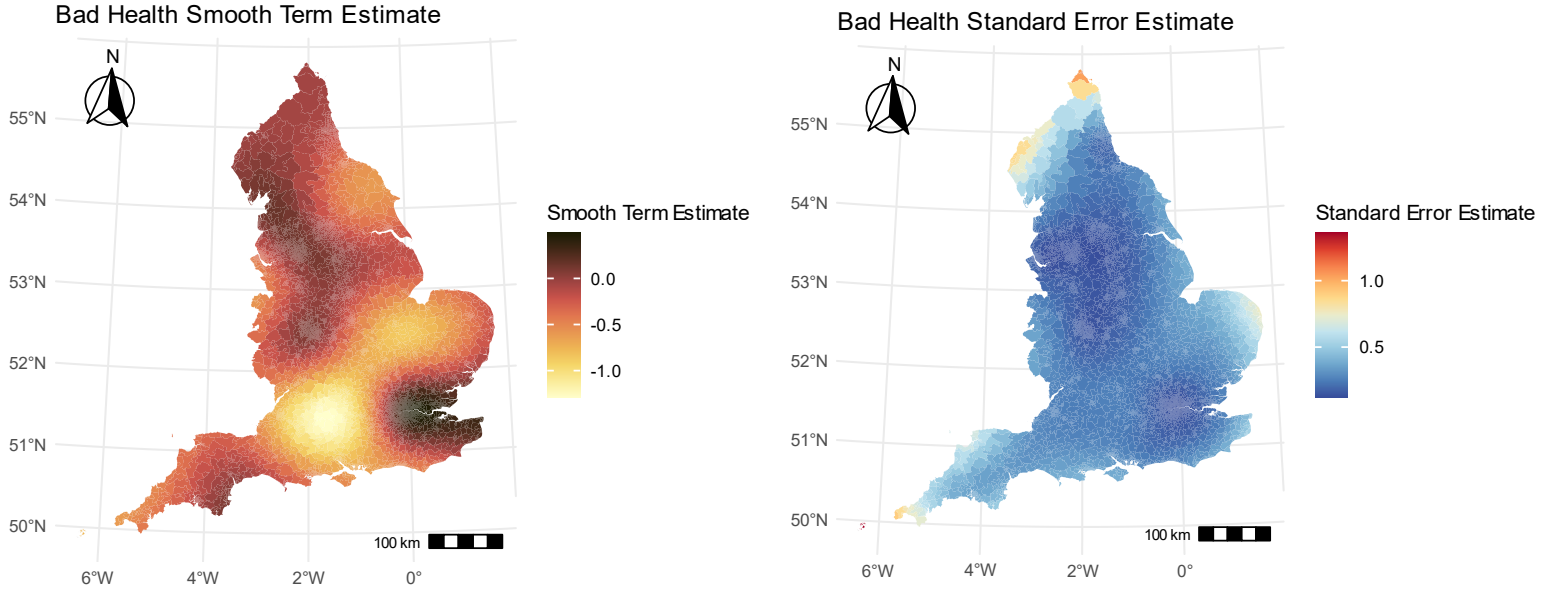

Figure 10: Maps showing the smooth term estimate and standard error for the Bad Health Variable.

The bad health variable has a much greater difference in smooth term estimates in comparison to the other variables, with London and the South-East seeing a higher rate of

increase in vaccination rate with an increase in the % of those with bad health. The South-West, however, sees a decrease. A higher rate of standard error is also observed, particularly focused around the Scottish Border.

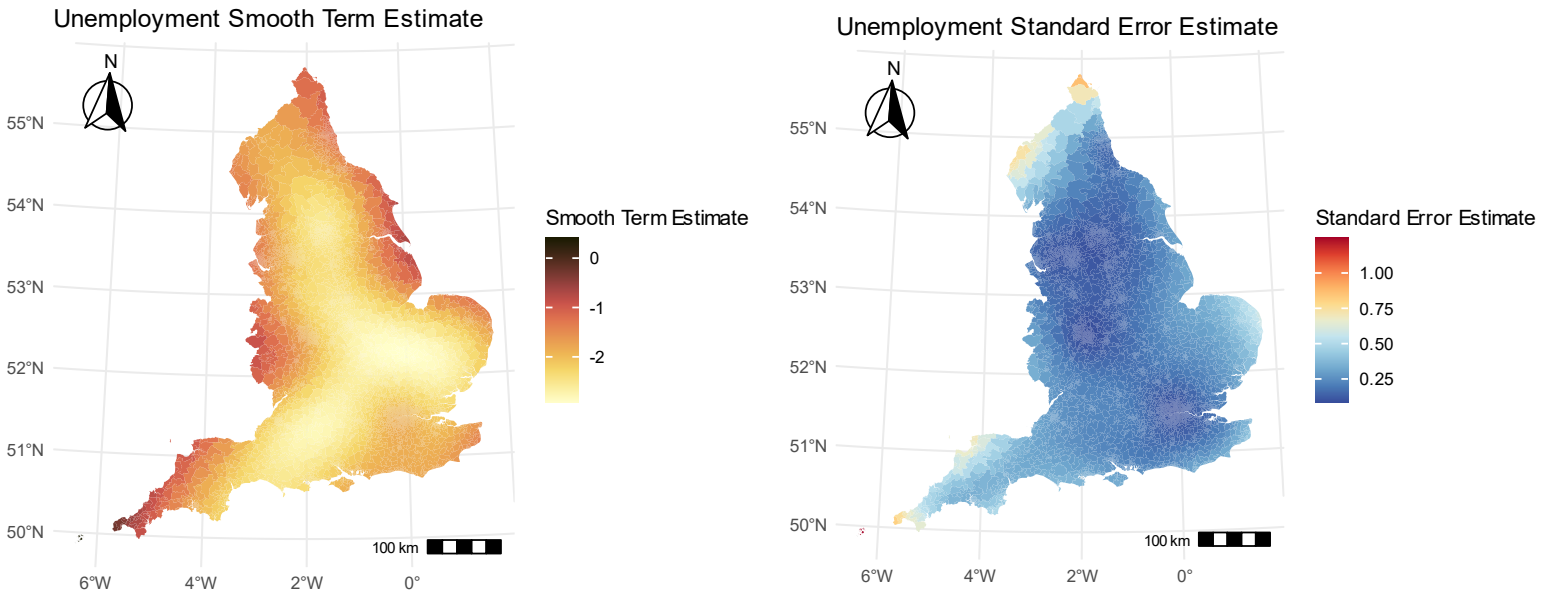

Figure 11: Maps showing the smooth term estimate and standard error for the Unemployment Variable.

Figure 11 suggests a greater rate of unemployment leads to a strong decrease in the vaccination rate. Although this effect is weaker in coastal regions and around England's borders. A higher rate of standard error is observed in England's northernmost MSOAs.

4. Discussion

4.1 Model Interpretation

The model created in this analysis exhibits strong explanatory power, evidenced by an adjusted R-squared of 0.917 and the ability to explain 99.7% of the deviance. Analysing the parametric coefficients revealed that only the intercept was globally significant. Visualisations aid in the recognition of the variables effects on the model locally, with an increase in the % over 65 and the % with level 4 qualifications being associated with an increase in vaccination rates, with the inverse relationship being true for the % in bad health and % unemployed.

The results of this GGP-GAM suggest that it is possible to accurately predict the vaccination rate in the UK using 4 key predictor variables and a tuned model which provides an understanding of the variables' non-linear relationships.

4.2 Critical Reflection on Model Choice

The flexibility of GAMs allows them to fit complex non-linear relationships in data, this is widely recognised as the major benefit of their usage within literature relating to spatial modelling (Comber et al., 2023a; Rigby and Stasinopoulos, 2005). The ability to produce smooth term estimates is particularly important in this context considering all variables registered high levels of significance locally and only the intercept globally. These local relationships would have gone unrecognised if a simple global regression model such as an Ordinary Least Squares (OLS) was selected.

The recognition of spatial variation in coefficients however is not exclusive to GAMs with Geographically Weighted Regression (GWR) models also providing the ability to investigate local trends in nonstationarity (Brunsdon et al., 1996). However, since the creation of the GWR, a Multiscale Geographically Weighted Regression (MGWR) has been created which can produce clearer local effects on parameter estimates (Fotheringham et al., 2017). MGWR is currently the brand-leader in the world of SVCs and is therefore the most important to compare when justifying the choice of a GAM.

GAMs have a relatively mature framework in comparison to the Geographically Weighted (GW) framework which has only been partially supported by package development in the major coding environments used in spatial analysis (Comber et al., 2023a, Comber et al. 2022). This means that issues such as overfitting are less likely in GAMs due to the penalised model options, developed to handle collinearity (Comber et al., 2023a).

The ability to handle collinearity combined with robust options for dealing with outliers and the non-constant variability of residuals across locations (heteroskedasticity) also provide a ‘more comprehensive theoretical underpinning’ for GAMs when compared to MGWR (Comber et al., 2023a, p.17).

The choice of a Gaussian Process (GP) as the basis of the GAM in this analysis is justified by research carried out by Comber et al. (2023b), in which GP splines were found to be effective in modelling complex spatial processes. This is because the use of GP splines allows for flexible, non-linear functions that can vary across space (Comber et al., 2023b). The analysis of Comber et al. (2023b) primarily compares a GGP-GAM and a MGWR and concludes that the GGP-GAM outperformed a GWR with more accurate coefficient estimates and lower residuals. Comber et al. (2023b, p.1) theorise that this difference could be attributed to an MGWR's inability to handle ‘highly localised patterns of spatial

heterogeneity'. This flaw in the MGWR could have been critical within the context of this research in which localised patterns were found. It is also recognised within this research and wider literature that GAMs display greater computation efficiencies than MGWRs which are computationally expensive (Comber et al., 2023a; Fan and Huang, 2022). Critical reflection compares a GGP-GAM approach favourably to MGWRs due to computational efficiency, stronger theoretical underpinning framework maturity strongly supporting the use of a GGP-GAM in this research.

4.3 Model Tuning

The model tuning undertaken in this analysis focused on adapting the `k` parameter. The choice of the `k` parameter in a GAM is not critical however it is important, it sets the upper limit on the degrees of freedom associated with the smooth and therefore the choice of `k` must be evaluated (Wood, 2017). The parameter selection `k` must be high enough to allow for a sufficient number of degrees of freedom which represent the underlying truth in the data, however, a high `k` choice can result in computational inefficiencies (Wood, 2017).

The `gam.check` informed that the method selected for the smoothing parameter was Generalised Cross-Validation (GCV), and the model produced a Root Means Squared GCV score gradient of 3.999294e-06 at convergence indicating that a good model fit for the data with relatively low complexity.

With the choice of `k` = 170, the model was able to converge after 4 iterations and had a positive definite Hessian indicating stability in the optimisation process.

	k'	edf	k-index	p-value
s(X,Y):Intercept	169.0	93.6	0.72	<2e-16 ***
s(X,Y):o65	170.0	33.9	0.72	<2e-16 ***
s(X,Y):l4qual	170.0	32.0	0.72	<2e-16 ***
s(X,Y):badhealth	170.0	19.7	0.72	<2e-16 ***
s(X,Y):unemp	170.0	21.4	0.72	<2e-16 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Figure 12: Table produced by `gam.check`

The k-index of 0.72 seen in Figure 12 is somewhat low, however diminishing returns were seen with a greater k value. The model rank of 277/854 suggests that the model is using a sufficient yet not exhaustive number of effective degrees of freedom indicating an appropriate level of complexity. The low p-values (<2e-16) highlight that the EDFs are not close to the maximum suggesting the model is not overfitting.

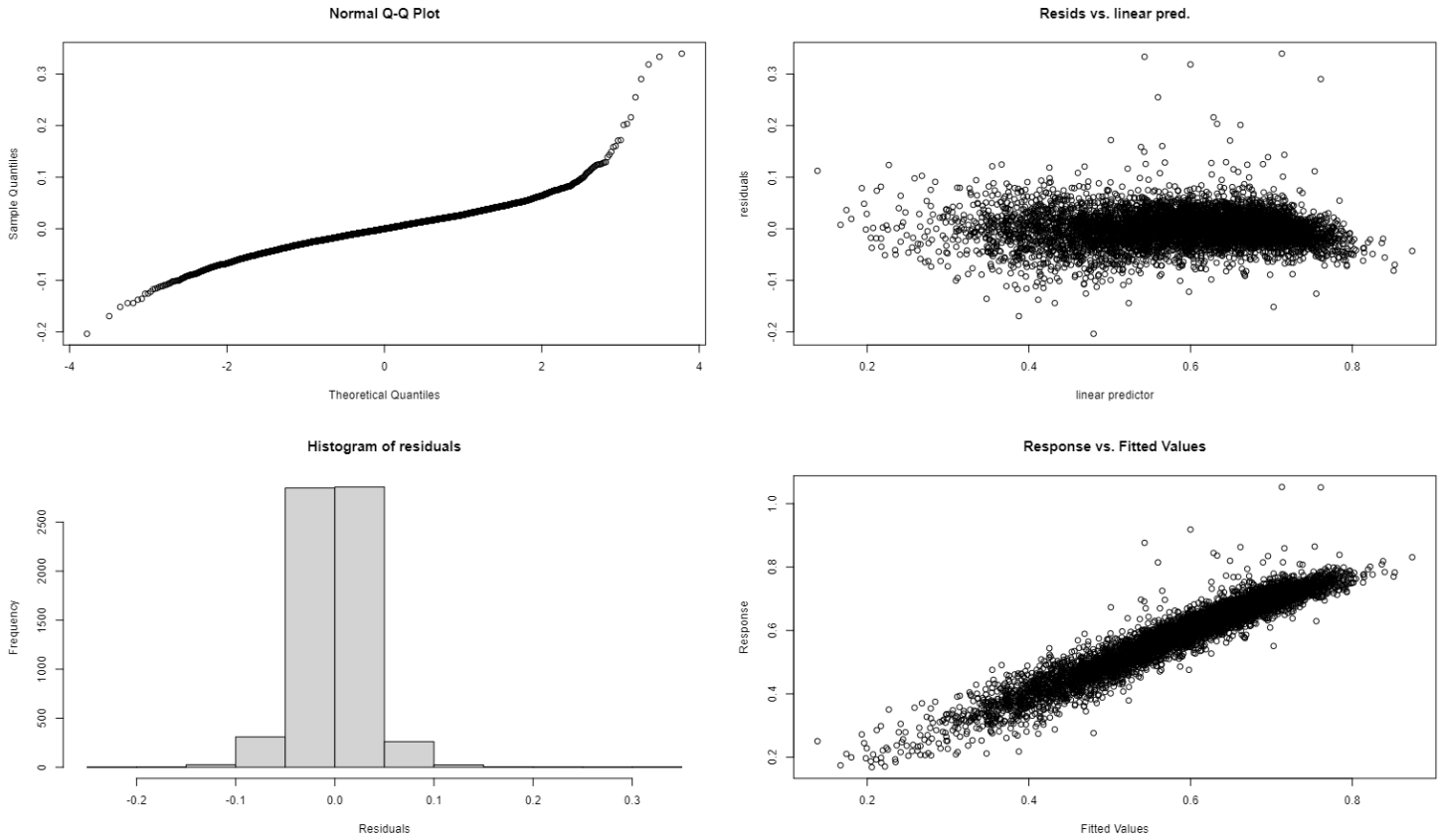

Figure 13: Plots produced by `gam.check`

An analysis of Figure 13 highlights the distribution of the residuals close to 0 in the histogram, combined with the clustering around the 1-to-1 line of the response vs. fitted values graph, suggesting a good model fit.

4.4 Limitations and Future Work

This analysis was conducted using data from England only. Any further research could include vaccination data for Wales and Scotland. This may help to explain variations in the data close to the borders and reduce the rate of standard error for variables observed close to the Scottish border.

This analysis only considered a small number of variables, other metrics such as mental health have however been recognised as significant factors and could be included in any further analysis (Wong et al., 2022).

GGP-GAMs are a relatively new concept that have not received as much coverage in literature. Further research and usage of GGP-GAMs could be useful in directing their construction and effectiveness in the future.

4.5 Conclusion

This paper demonstrates the ability of a GGP-GAM to model the complex and spatially varying relationships between socio-demographic variables and COVID-19 vaccination rates in England, outlined by a high adjusted R-squared and through analysis of the coefficients. The choice of a GGP-GAM was crucial in understanding the non-linear relationships and spatial heterogeneity in the data as highlighted by local significance for all predictor variables. Whilst acknowledging the successes of this analysis it is important to recognise the limitations and how the inclusion of data for Scotland and Wales could provide further insights in the future, along with greater insight from literature into the construction and benefits of GGP-GAMs.

Appendix

1. **Reproducibility statement**

   All data and code used in this assignment along with high-resolution copies of the maps can be found in GitHub here (https://github.com/freddie-wallace/GEOG3195_Assignment).

References

Brunsdon, C., Fotheringham, A.S. and Charlton, M.E. 1996. Geographically weighted regression: A method for exploring spatial nonstationarity. Geographical analysis. 28(4), pp.281–298.

Comber, A., Callaghan, M., Harris, P., Lu, B., Malleson, N. and Brunsdon, C. 2022. Gwverse: A template for a new generic geographically weighted R package. Geographical analysis. 54(3), pp.685–709.

Comber A, Harris P, Murakami D, Nakaya T, Tsutsumida N, Yoshida T and Brunsdon C. 2023a. Spatially varying coefficient modelling with a Geographical Gaussian Process GAM (GGP-GAM). Paper submitted to Geographical Analysis (April 2023).

Comber, A., Harris, P. and Brunsdon, C. 2023b. Multiscale spatially varying coefficient modelling using a Geographical Gaussian Process GAM. Geographical Information Systems., pp.1–21.

Fan, Y.-T. and Huang, H.-C. 2022. Spatially varying coefficient models using reduced-rank thin-plate splines. Spatial statistics. 51(100654), p.100654.

Fotheringham, A.S., Yang, W. and Kang, W. 2017. Multiscale geographically weighted regression (MGWR). Annals of the American Association of Geographers. 107(6), pp.1247–1265.

Rigby, R.A. and Stasinopoulos, D.M. 2005. Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society. Series C, Applied statistics. 54(3), pp.507–554.

Wong, C.L., Leung, A.W.Y., Chung, O.M.H. and Chien, W.T. 2022. Factors influencing COVID-19 vaccination uptake among community members in Hong Kong: a cross-sectional online survey. BMJ open. 12(2), p.e058416.

Wood, S.N. 2017. Generalized additive models: An introduction with R, second edition. Boca Raton, FL: CRC Press.