

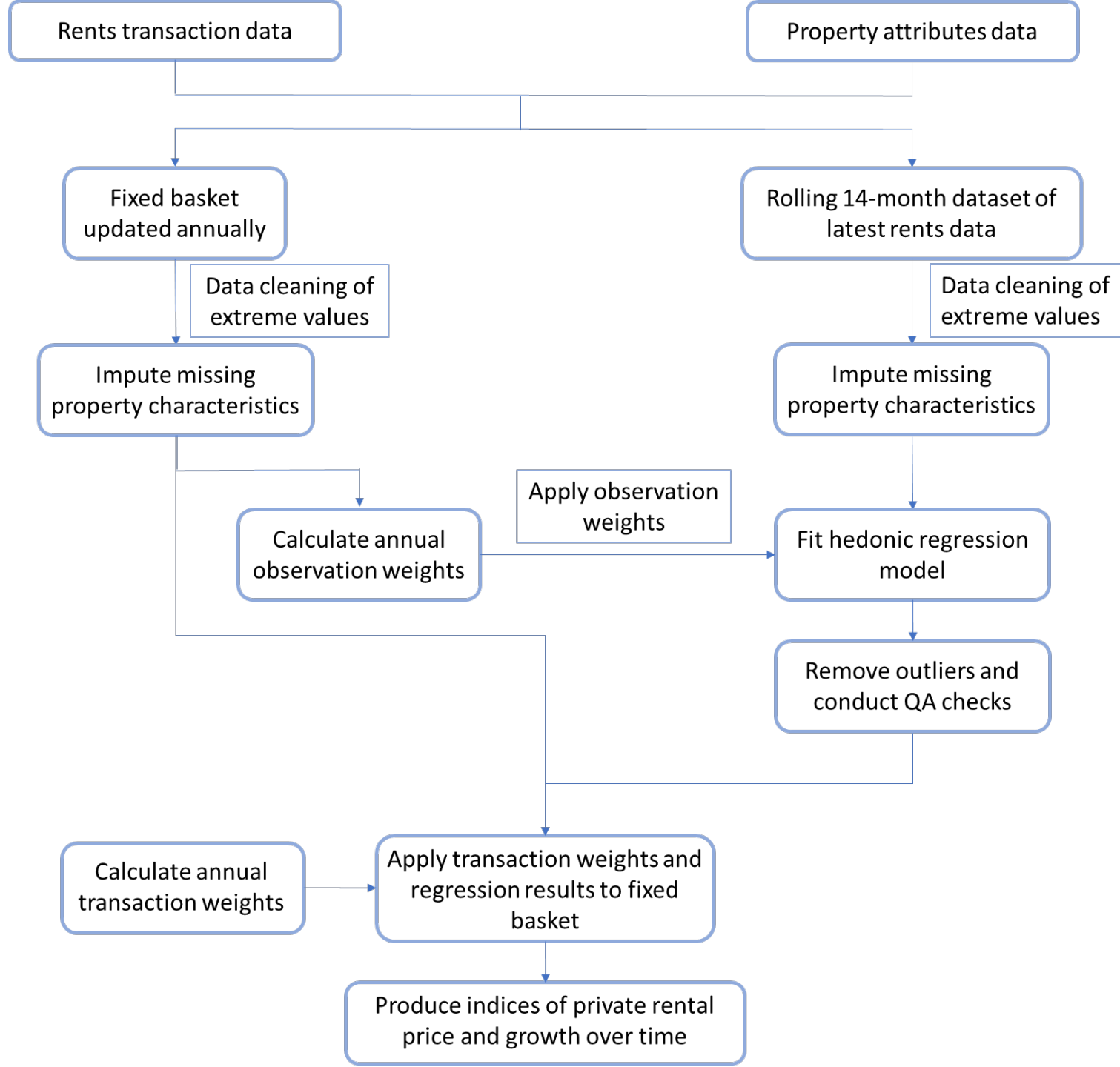
OFF-SEN: Transformation of private rental price
statistics:
Proposed hedonic regression models

Current method (IPHRP)

- ‘Matched pairs’ method heavily weighted towards a ‘stock’ measure
- Each January, all data collected within the last 14 months is split 50:50 into a sample pool and a substitution pool
- New data collected each month is compared with the sample
 - If new data for a property in the sample is collected, the price is updated
 - For properties in the sample which have no price update in the latest month, price is carried forward
 - Properties in the latest month’s data that are not in the sample pool are sent to the substitution pool
 - Properties in the sample that have not had a price update for 14 months are removed and a comparable replacement is selected from the substitution pool
- Produces index down to region level only

Introduction

Project aim	To use transaction-level price microdata to transform private rental price statistics
Anticipated outputs for publication	Index, annual rental growth and average rental price timeseries at the following levels: • UK • Countries • English regions • Local authorities (3-month average, as published by HPI) Breakdowns at by: • Property type (detached, semi-detached, terraced, flat) • Bedroom category (studio, 1-bedroom, 2-bedroom, 3-bedroom, 4+ bedroom)
Methodology research	• UK House Price Index (HPI) • UK Index of Private Housing Rental Prices (IPHRP) • England Private Rental Market Summary Statistics (PRMS)



Property characteristics used to predict rental price

Property characteristic	Source	% categories that are statistically significant ($p < 0.05$)
Number of bedrooms	VOA Council Tax data	100%
Floor area	VOA Council Tax data	100%
Property age	VOA Council Tax data	100%
Local authority	GeoPortal Postcode lookup	98%
Acorn group	ACORN CACI data	100%
Property type	Price microdata (VOA, Welsh Government)	100%
Furnished status	Price microdata (VOA, Welsh Government)	100%

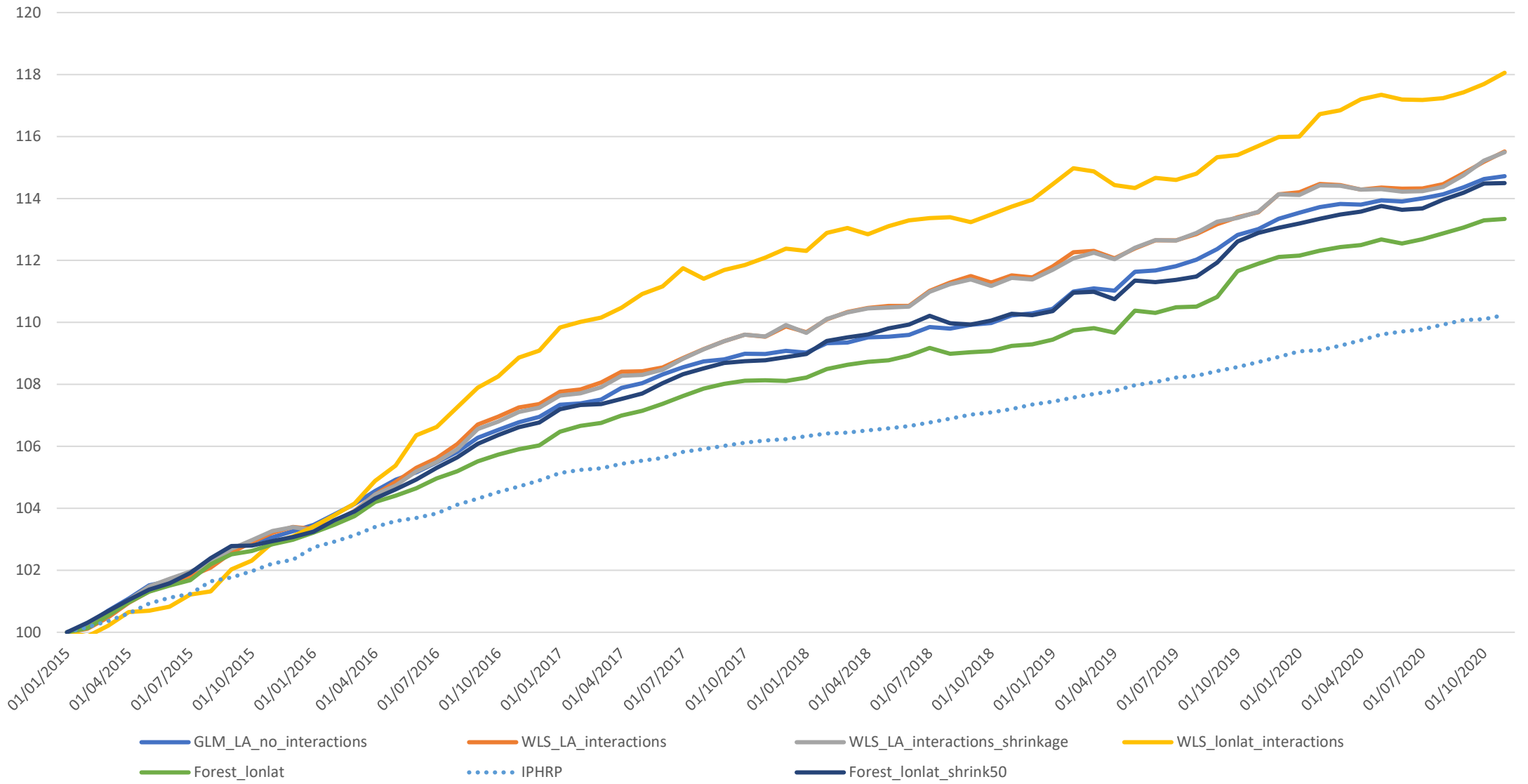
Random forest regression vs general linear model

- GLMs assume there is a linear relationship between price and explanatory variables, whereas random forests do not make this assumption
- Random forests account for non-linearities and interactions between explanatory variables automatically, whereas the analyst must specify this for GLM
- Random forests have many ways to fine-tune (for better predictive performance) and regularise (to prevent overfitting) the model
- Random forests are more of a “black box”. GLM outputs regression coefficients which enable analysts to assess the contribution of each variable to the price and better explain a change in the index

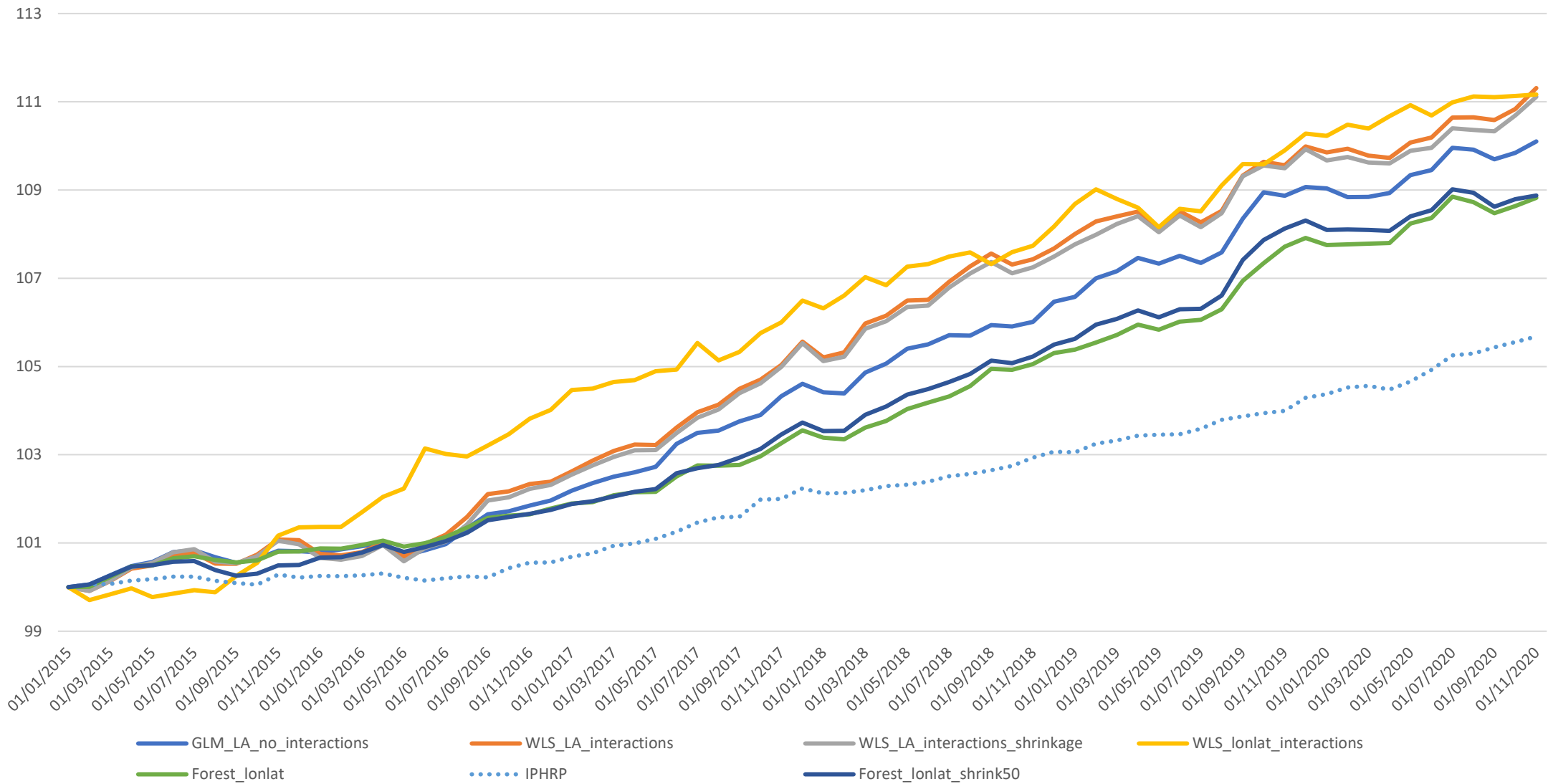
Overview

- Last month, ONS presented results from two regression models:
 - WLS without interaction terms
 - Random Forest
- Random forest results looked reasonable, and showed a similar trend to WLS at region and country level
- However, some local authorities showed unrealistic results at LA-level
- APCP-T suggested application of shrinkage to prevent over-fitting and reduce volatility
- Consequently, ONS tested four additional regression models, including the impact of shrinkage

England

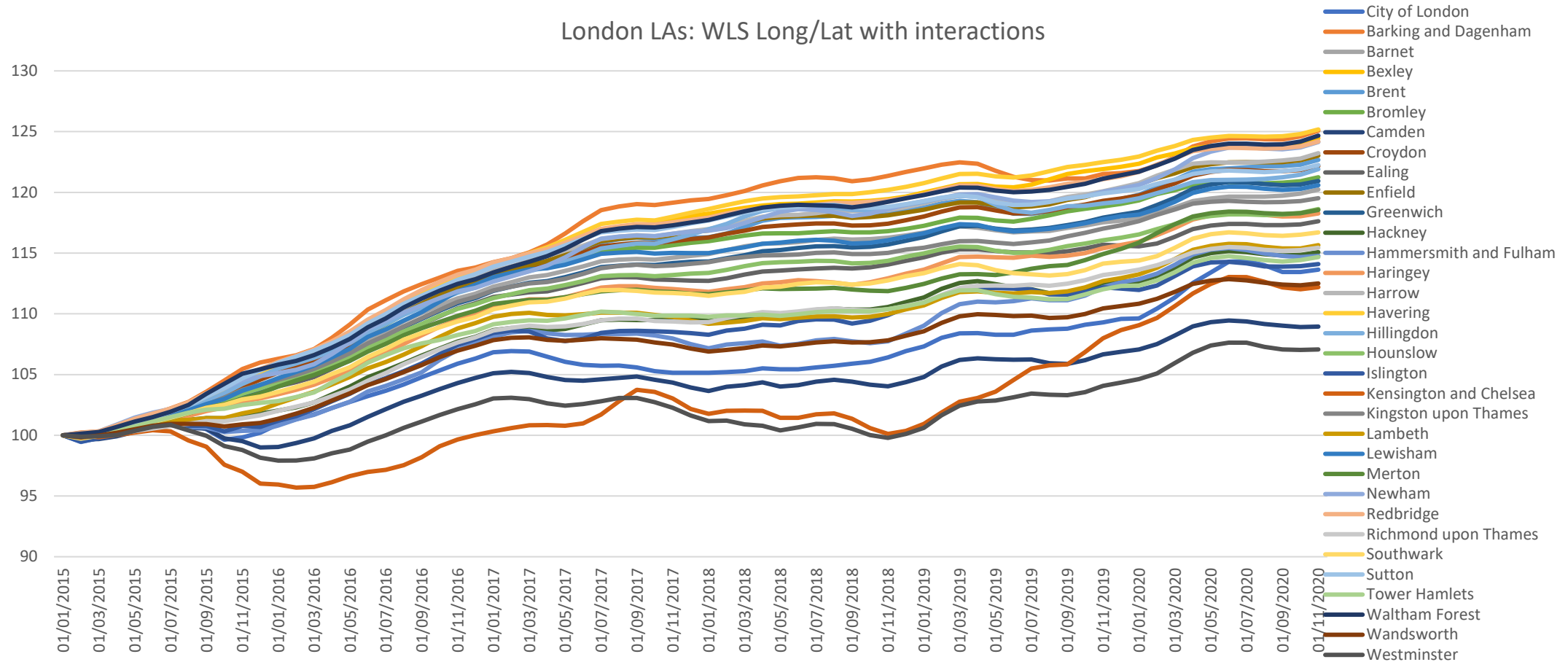


Wales

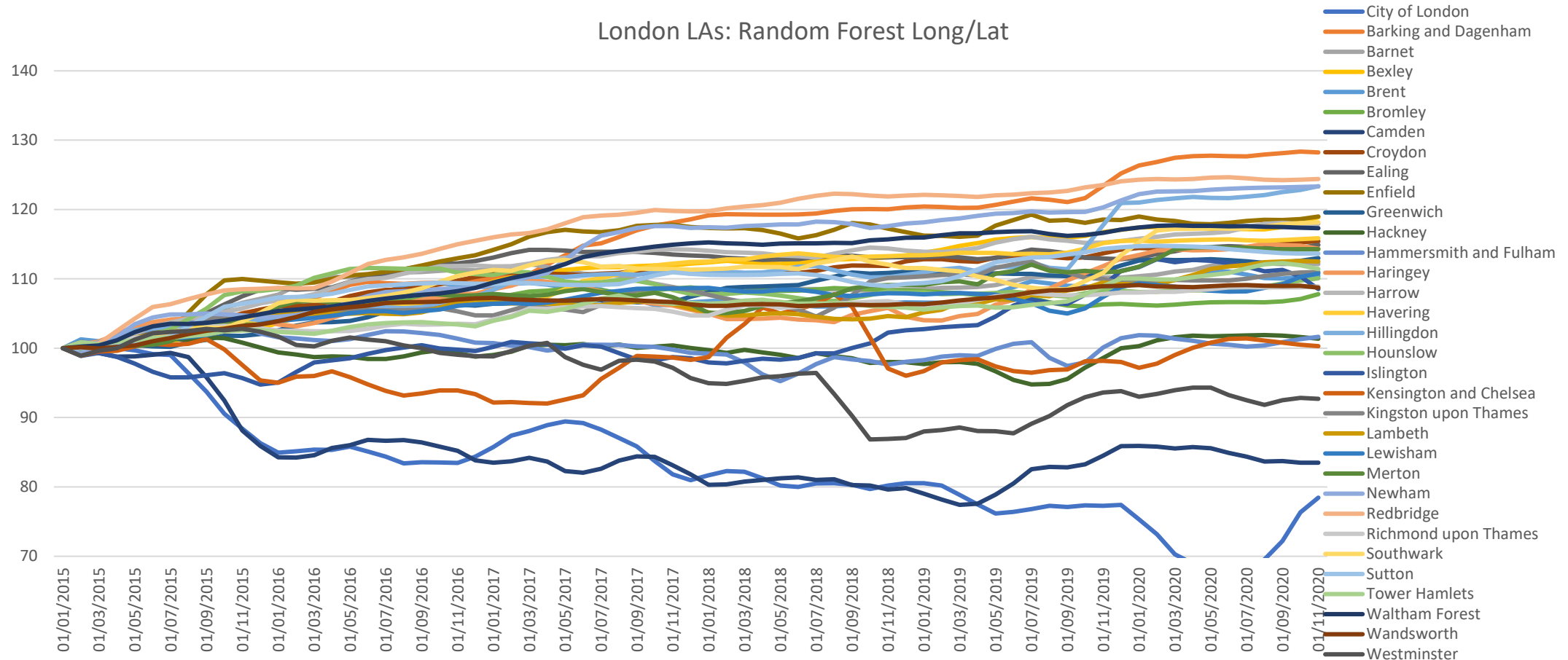


Rejected models

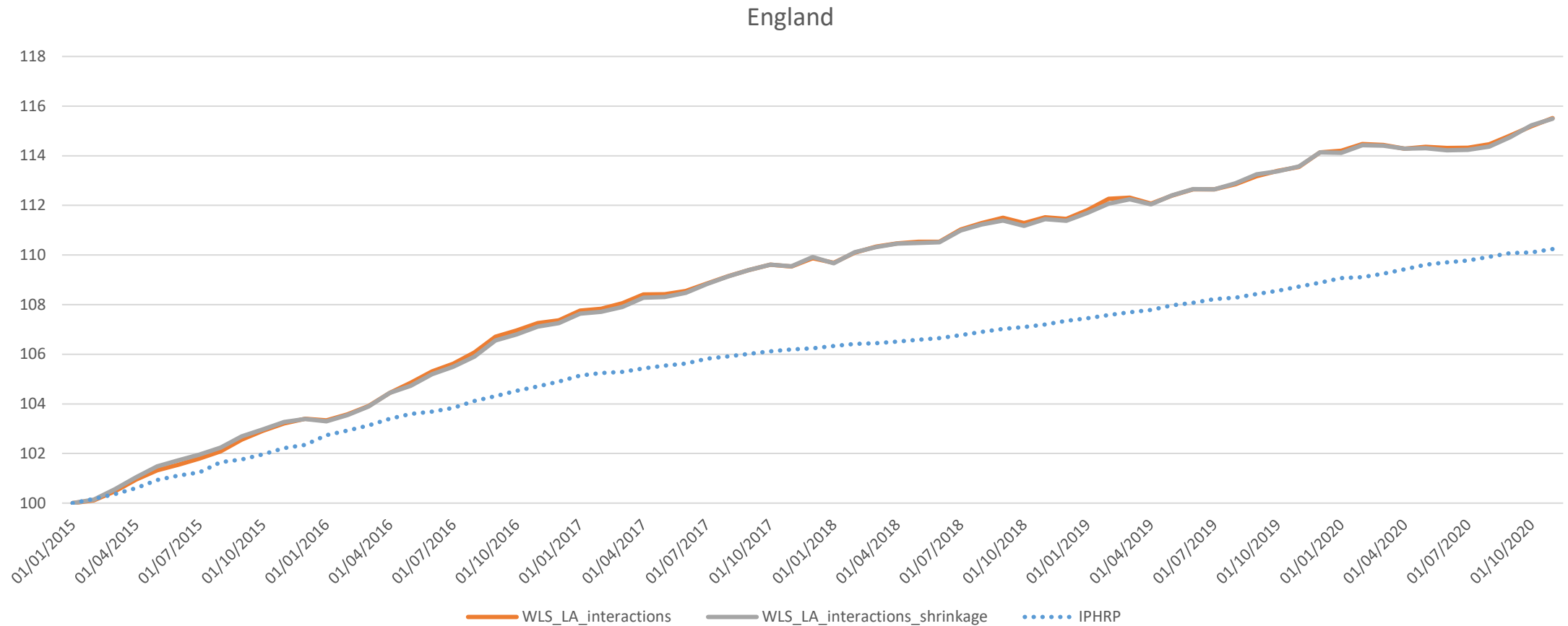
WLS using longitude/latitude instead of LA code with interactions (acorn vs area, type, beds)



Random Forest using longitude/latitude, no shrinkage



WLS using LA with interactions (acorn vs area, type, beds) and shrinkage



Decision: Eliminate these three model options

WLS with longitude/latitude is too simplistic/unrealistic as it models LAs in close proximity the same way

Random Forest without shrinkage is too volatile/unrealistic at LA-level

WLS with shrinkage is almost identical to the model without shrinkage. Without shrinkage model offers more statistical information

Regression models for
consideration

Regression models for consideration

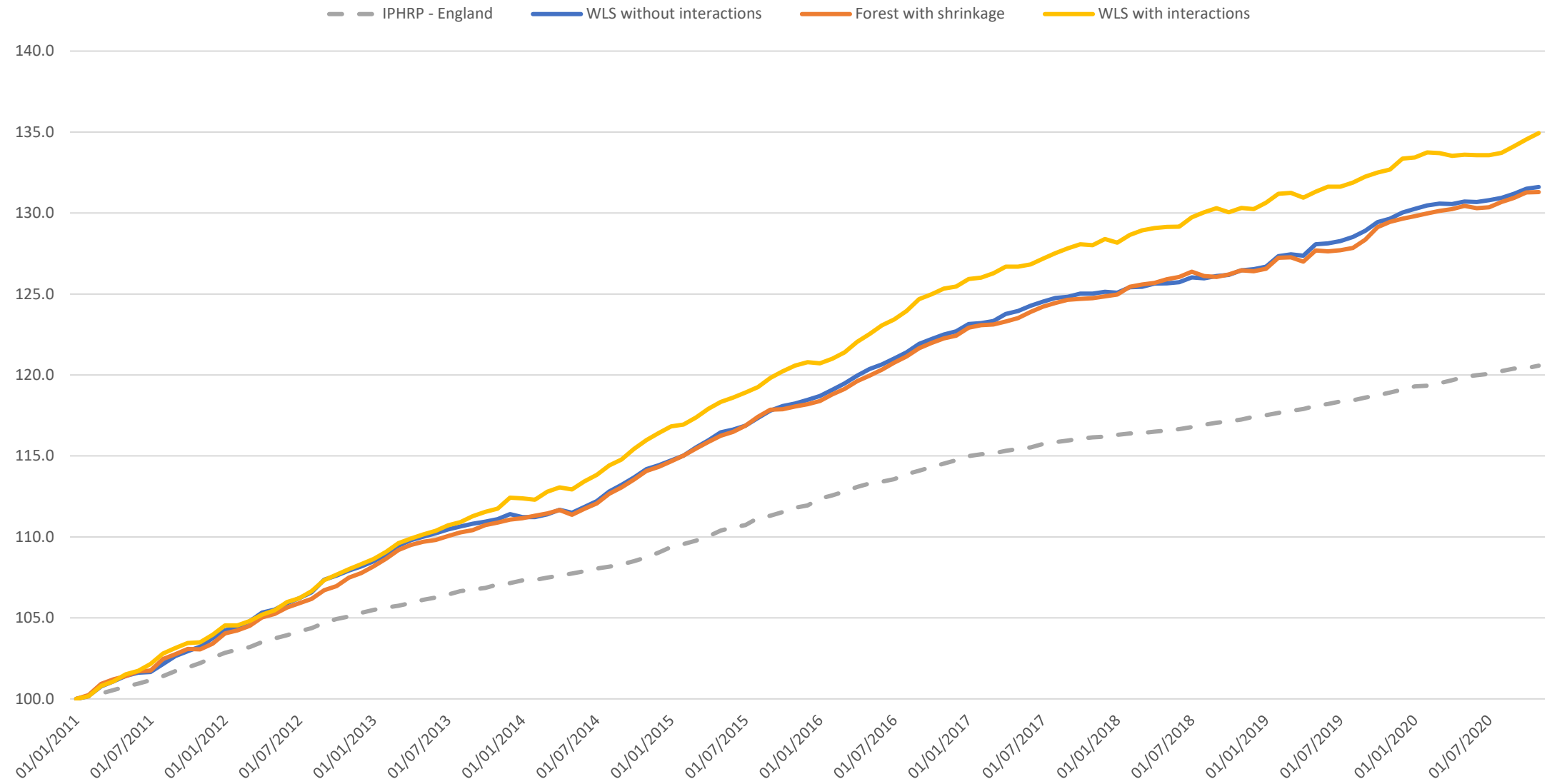
- Model 1 – Weighted Least Squares without interaction terms
- Model 2 – Random Forest with shrinkage
- Model 3 – Weighted Least Squares with some interaction terms

Preferred model:

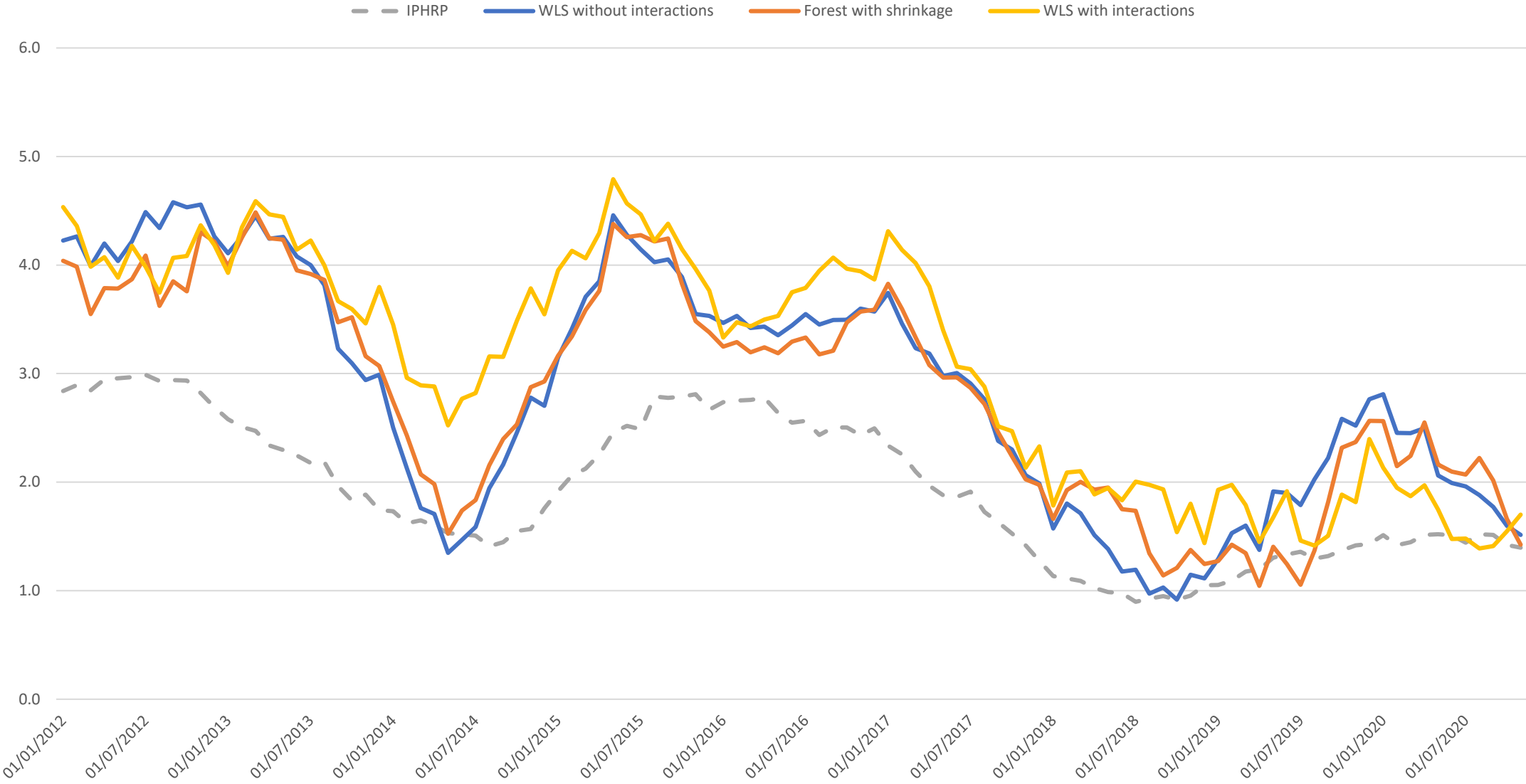
Model 1 – WLS without interactions

- ONS Regression experts recommend random forest (Model 2) as the most suitable model in practice
- However, ONS must balance theory with practicalities: the model must be suitable for use in regular production
- WLS is a more transparent method than random forest – easy to explain to users and interpret model outputs
- Model 1 agrees very closely with the random forest
- ONS is concerned Model 3 might not account for all interactions and produces very different results compared with Models 1 & 2
- Model 1 is easier to quality assure data and results than Model 2

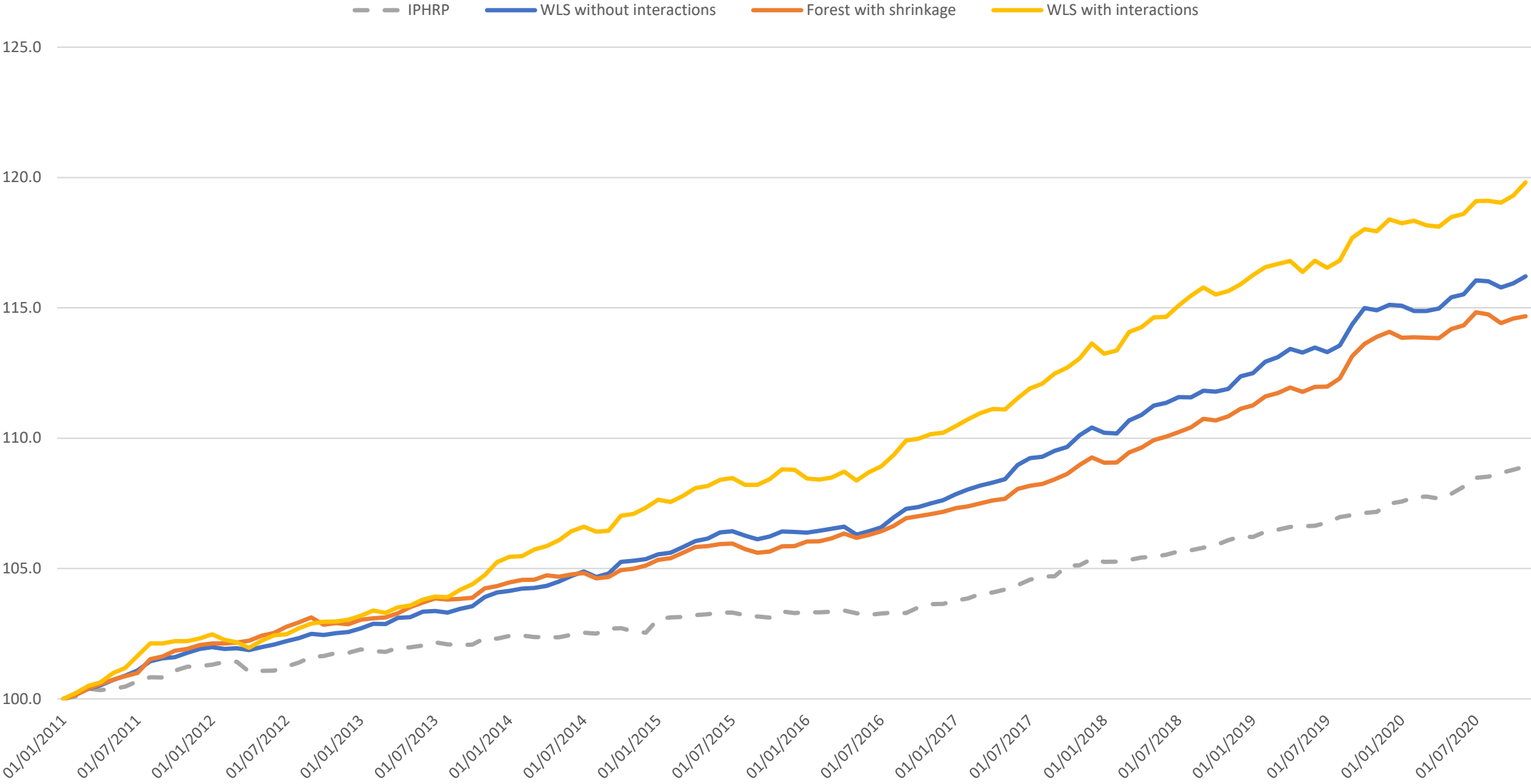
England index



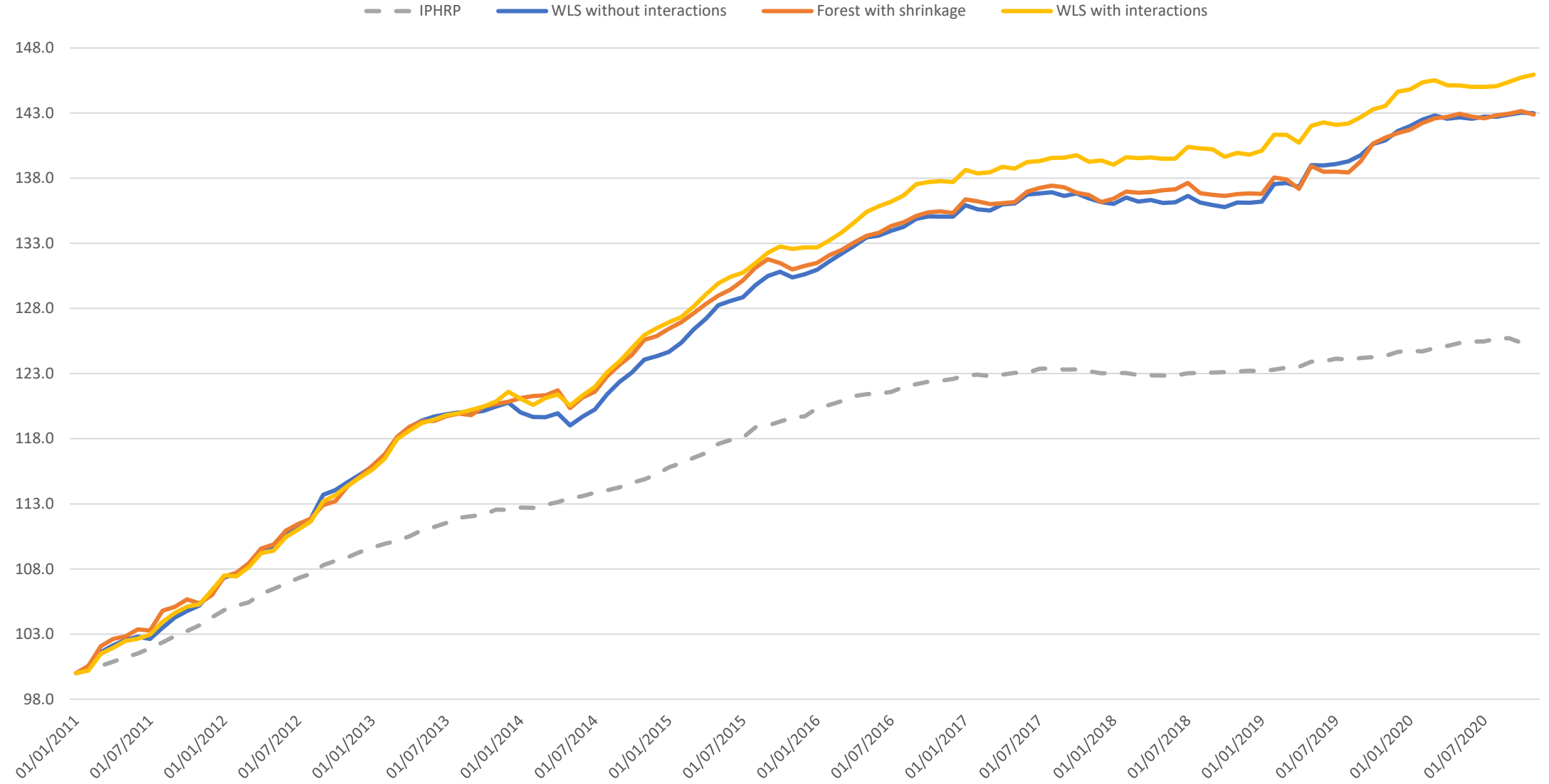
England annual % growth



Wales index



London index



Next steps

- Select preferred regression model
 - Originally a general linear model was proposed
 - ONS regression methodology experts recommended switching to a random forest regression model to better account for interactions between variables
 - Six models have been tested; three have been rejected
 - Gather feedback from APCP-T on the three remaining proposed models
- Ratify methodology through ONS Methodology Advisory Committee
- Initiate user consultation period: Spring-Summer 2021
- Publish initial analysis & collect user feedback: Summer-Autumn 2021

Do you have any comments on
the three regression models?
Which model is your preference?