Government Statistical Service

Preliminary Research into Internet Data Sources

Susan Williams and Martin Ralphs

Standards & Strategy Division

Office for National Statistics

26th June 2013

Topics

- Context: internet tools research at ONS
- Internet research topics in Beyond 2011
- Case Study: Google Insights and Migration

Internet tools research at ONS

- Internet search tools may be able to help ONS in several ways:
  - By providing quality assurance around other statistics – do patterns look the same?
  - By providing an indicator about possible change on the ground – e.g. Influx of new visitors to the UK, dispersal of migrant populations.
  - By informing statistical processes, e.g. direct inputs into models or to operational processes. This is a LONG WAY OFF!

Context: “Beyond 2011” Programme

- Identify the best way to provide small area population and socio-demographic statistics in future

- Provide a recommendation in September 2014
  - underpinned by full cost-benefit analysis
  - & high level design for implementation

- (subject to agreement) Implement the recommendation

- http://www.ons.gov.uk/beyond2011

Context: the caveats

- This is **VERY PRELIMINARY WORK**
- Very much an initial exploration of potential and pitfalls
- The results we have observed are interesting, but there are complex issues around meaningful interpretation
- A lot more work is needed before we can make any recommendations about the utility and robustness of Google Insights data
- **A long way from a “magic bullet”**

Our interest...

- What sort of information might we extract from internet transaction data like Google searches?
- Preliminary studies (e.g. Google Flu) suggest that these data can provide insight into emerging events
- Could we harness these data to validate other information, to flag potential events of interest and / or to improve the quality of socio-economic statistics?

Case Study: Google Insights

- The Internet is used by a large proportion of the UK population – smart phones/tablets;
- Search engines are used by most people
- The main search engine in the UK is Google
- Google have created an application to analyse the searches made through Google
- Google Insights for Search www.google.com/insights/search/ - recently merged into www.google.com/trends/

Google Insights – basics

- Based on a **sample** of search queries (which changes each day) - sample size is not straightforward to obtain
- Weekly data (**usually**)
- Search Volume Index (SVI) – **not** actual transaction volumes!
  - Weekly proportions: Volume of searches of interest / All searches (within geographic area)
  - The week with the largest proportion is set to 100
  - All other proportions are scaled back appropriately

Google Insights for Search beta

Compare by
- Search terms
- Locations
- Time Ranges

Search terms

Tip: Use a comma as shorthand to add comparison items. (tennis, squash)

All search terms

+ Add search term

Filter

Web Search | United Kingdom | All sub-regions | 2004 - present | All Categories | Search

Web Search Interest

United Kingdom, 2004 - present

The categorisation taxonomy of Google Insights for Search has been updated during December 2011. Learn more

An improvement to our geographical assignment was applied retroactively from 1/1/2011. Learn more

Search terms

Top searches

1.	facebook	
2.	bbc	
3.	youtube	
4.	hotmail	
5.	ebay	
6.	you	
7.	google	
8.	games	
9.	news	
10.	weather	

Google Embed this table

Rising searches

1.	facebook	Breakout
2.	facebook login	Breakout
3.	youtube	Breakout
4.	google	+500%
5.	hotmail	+300%
6.	you	+200%
7.	mail	+170%
8.	ebay	+130%
9.	bbc	+110%
10.	amazon	+100%

Google Embed this table

Google Insights for Search beta

Compare by
- Search terms
- Locations
- Time Ranges

Search terms

Tip: Use the minus sign to exclude terms (wimbledon -tennis).

All search terms

+ Add search term

Filter

Web Search

United Kingdom | All sub-regions

2004 - present

TV & Video Equipment

Search

Web Search Interest

United Kingdom, 2004 - present

All Categories > Shopping > Consumer Electronics > TV & Video Equipment

The categorisation taxonomy of Google Insights for Search has been updated during December 2011. Learn more

An improvement to our geographical assignment was applied retroactively from 1/1/2011. Learn more

Interest over time

forecast | News headlines

How can I see numbers?

The last value prior to the forecast is based on partial data and may change. Learn more

Future values are based exclusively on the extrapolation of past values. Learn more

Google Embed this chart

Beyond 2011 application: in-migration

- EU expansion 2004 (EU8 + Malta/Cyprus) & 2007 (Bulgaria & Romania)
- Polish migrants the biggest EU8 population now in the UK: 500K+
- What can Google Insights tell us about national and sub-regional patterns of in-migration from EU8 countries?
- Will non-English speaking migrants be searching in Google in their native language??

http://www.google.com/insights/search/#q=polski&geo=GB&cmpl

Google Insights for Search - ...

Google Insights for Search beta

+1 Help | Sign in | Download as CSV | English (UK)

Compare by

- Search terms
- Locations
- Time Ranges

Search terms

Tip: Use quotation marks to match an exact phrase ("table tennis").

polski

+ Add search term

Filter

Web Search

United Kingdom | All sub-regions

2004 - present

All Categories

Search

Web Search Interest: polski

United Kingdom, 2004 - present

Categories: Reference (25-50%), Arts & Entertainment (0-10%), Law & Government (0-10%) , more...

The categorisation taxonomy of Google Insights for Search has been updated during December 2011. Learn more

An improvement to our geographical assignment was applied retroactively from 1/1/2011. Learn more

Interest over time

forecast | News headlines

How can I see numbers?

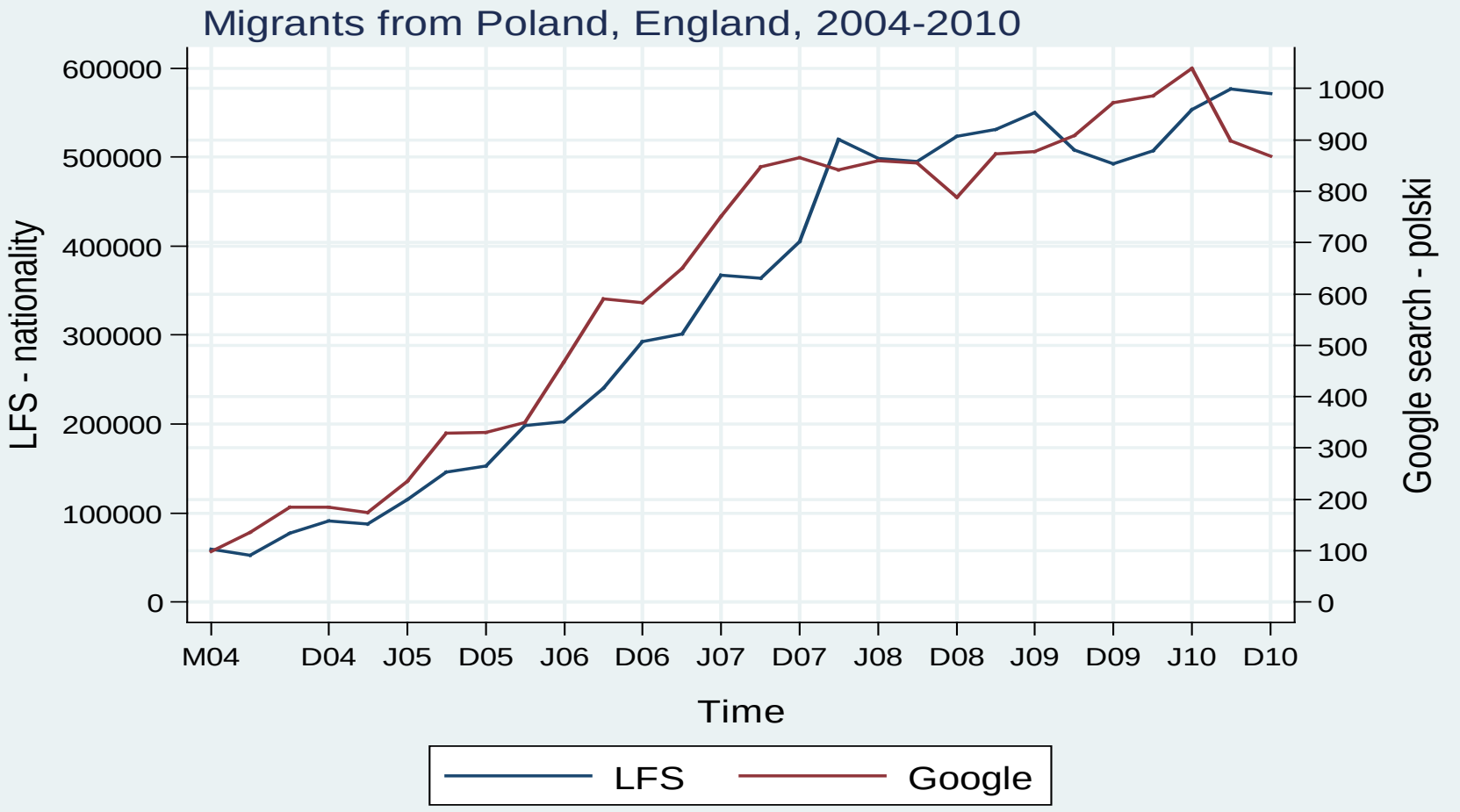
The last value on the graph is based on partial data and may change. Learn more

Google Embed this chart

Regional interest

Sub-region Town/City

Insights SVI compared to Labour Force Survey: Polish Nationals in England



Source: Labor Force Survey, Google Insight

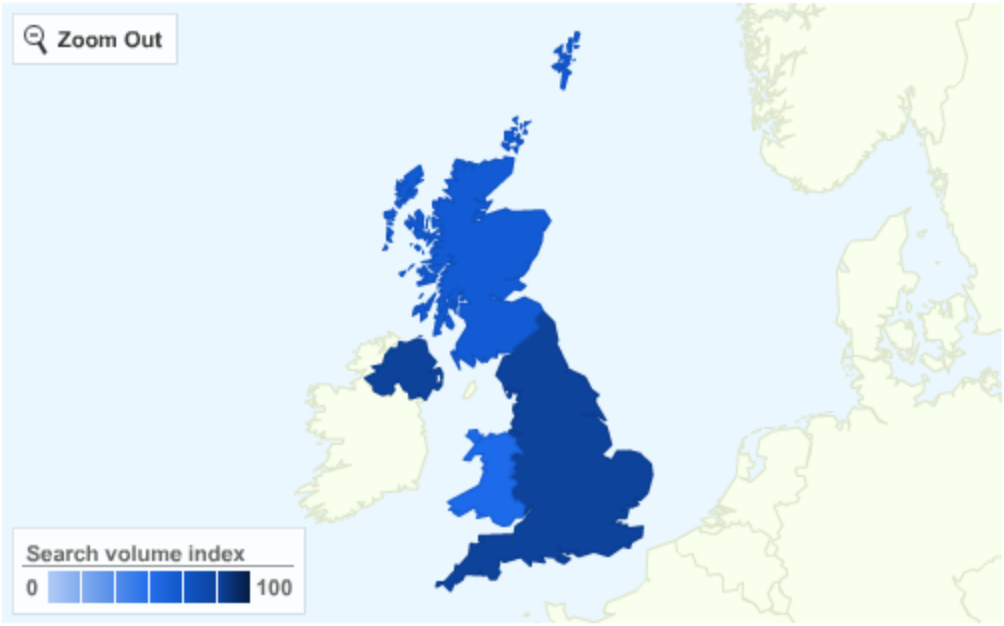
http://www.google.com/insights/search/#q=polski&geo=GB&cmpl

Google Insights for Search - ...

Regional interest

Sub-region Town/City

1. Northern Ireland	
2. England	
3. Scotland	
4. Wales	



View change over time

Search terms

Top searches

1.	angielsko polski	
2.	słownik	
3.	słownik polski	
4.	mapa	
5.	mapa polski	
6.	słownik angielsko polski	
7.	slownik	
8.	slownik polski	
9.	slownik angielsko polski	

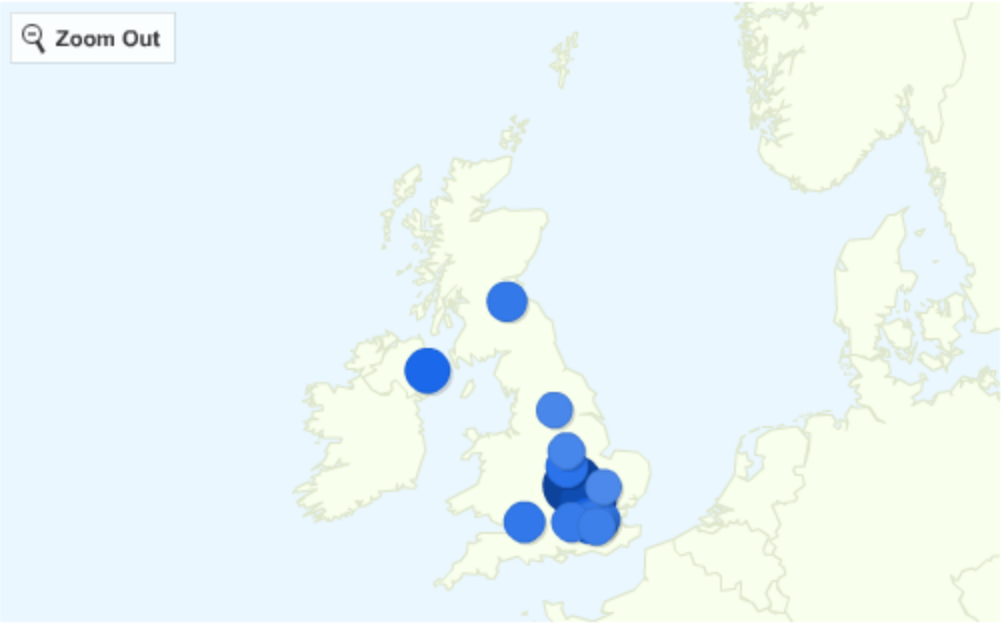
Rising searches

1.	polski sklep	Breakout
2.	słownik	Breakout
3.	słownik angielsko polski	Breakout
4.	słownik polski	Breakout
5.	translator angielsko polski	Breakout
6.	tłumacz	Breakout
7.	tłumacz angielsko polski	Breakout
8.	angielsko polski	+750%
9.	slownik angielsko polski	+170%

Regional interest

Sub-region Town/City

1.	Northampton
2.	Luton
3.	Brentford
4.	Belfast
5.	Leicester
6.	Bristol
7.	Poplar
8.	London
9.	Kensington
10.	Edinburgh



Search terms

Top searches

1.	angielsko polski
2.	słownik
3.	słownik polski
4.	mapa
5.	mapa polski
6.	słownik angielsko polski
7.	slownik
8.	slownik polski
9.	slownik angielsko polski
10.	tłumacz

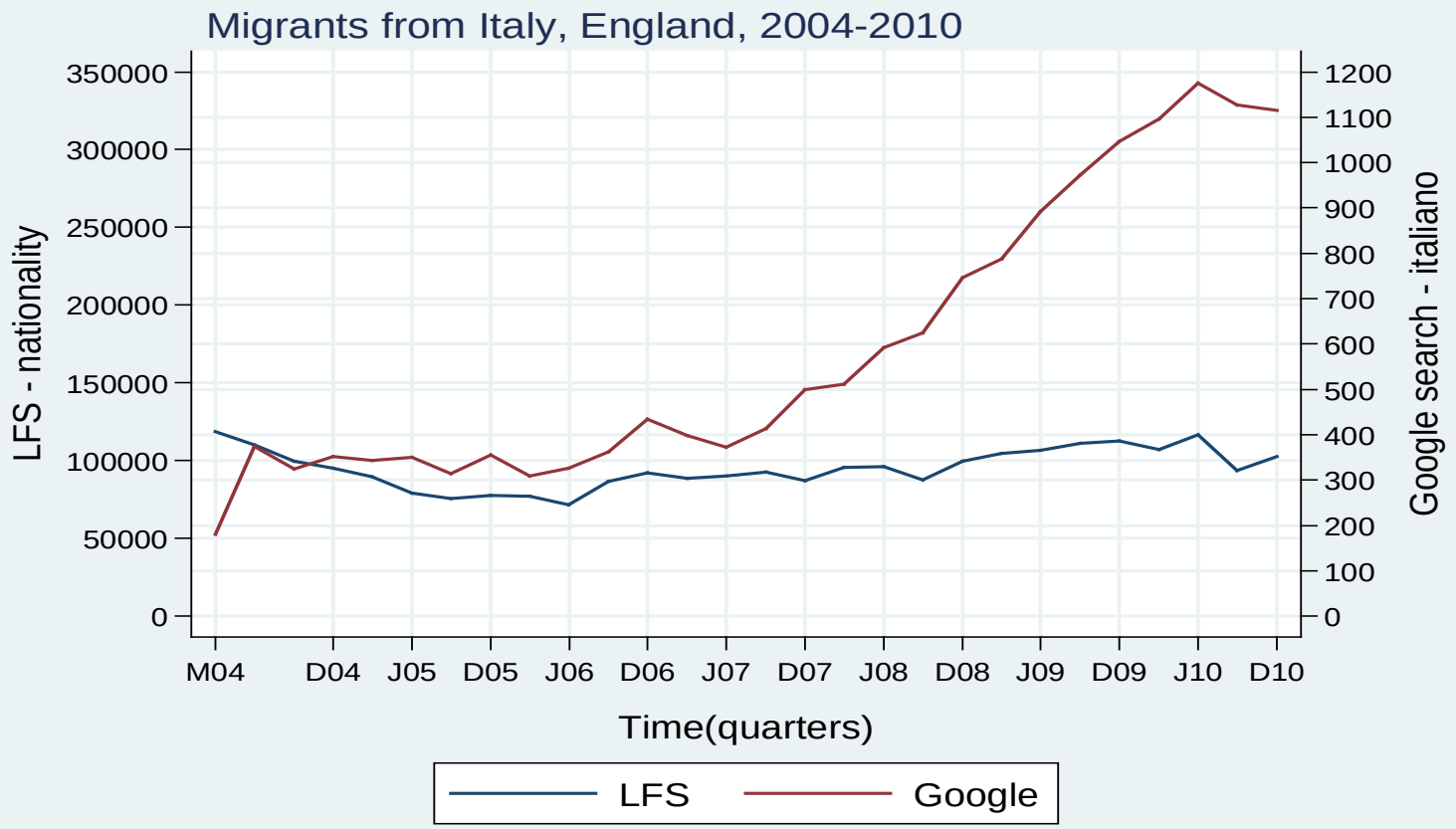
Rising searches

1.	polski sklep	Breakout
2.	słownik	Breakout
3.	słownik angielsko polski	Breakout
4.	słownik polski	Breakout
5.	translator angielsko polski	Breakout
6.	tłumacz	Breakout
7.	tłumacz angielsko polski	Breakout
8.	angielsko polski	+750%
9.	slownik angielsko polski	+170%
10.	slownik polski	+140%

Investigation on other nationalities within the UK

- Promising results for Lithuanians and Romanians
- Less success with smaller EU8 populations (less than 50K)
- More established populations such as Indians, Bangladeshis and Pakistanis more difficult (English commonly spoken, multiple languages; unable to differentiate from individuals with British nationality).
- Searches on “Italiano” showed diverging timeseries from LFS up to 1 Jan 2011 where there was a steep fall in the SVI:
  - Increased short-term migration? More interest in things “Italiano”?
  - Discontinuities (at 1 Jan 2011) - corresponding to methodological change on location?

Searches on “Italiano” in England



Research within rest of Europe

- Searches originating in other European countries : using “polski” term
- Large drops in SVI at 1 Jan 2011 (possibly due to methodological change on location) – esp. Italy and Spain
- Germany and Austria opened their borders to the EU8 on May 1 2011.

http://www.google.com/insights/search/#q=polski&geo=AT&cmpl

Google Insights for Search - ...

Google Insights for Search beta

+1 Help | Sign in | Download as CSV | English (UK)

Compare by

- Search terms
- Locations
- Time Ranges

Search terms

Tip: Use the plus sign to indicate OR (tennis + squash).

polski

+ Add search term

Filter

Web Search

Austria | All sub-regions

2004 - present

All Categories

Search

Web Search Interest: polski

ustria, 2004 - present

ategories: Reference (10-25%), Arts & Entertainment (0-10%), People & Society (0-10%) , more...

The categorisation taxonomy of Google Insights for Search has been updated during December 2011. Learn more

An improvement to our geographical assignment was applied retroactively from 1/1/2011. Learn more

nterest over time

forecast | News headlines

How can I see numbers?

The last value on the graph is based on partial data and may change. Learn more

Google Embed this chart

Regional interest

Sub-region Town/City

http://www.google.com/insights/search/#q=polski&geo=DE&cmpl

Google Insights for Search - ...

Google Insights for Search beta

+1 Help | Sign in | Download as CSV | English (UK)

Compare by

- Search terms
- Locations
- Time Ranges

Search terms

Tip: Use a comma as shorthand to add comparison items. (tennis, squash)

polski

+ Add search term

Filter

Web Search

Germany | All sub-regions

2004 - present

All Categories

Search

Web Search Interest: polski

ermany, 2004 - present

ategories: Reference (10-25%), Arts & Entertainment (0-10%), Business & Industrial (0-10%) , more...

The categorisation taxonomy of Google Insights for Search has been updated during December 2011. Learn more

An improvement to our geographical assignment was applied retroactively from 1/1/2011. Learn more

nterest over time

forecast | News headlines

How can I see numbers?

The last value prior to the forecast is based on partial data and may change. Learn more

Future values are based exclusively on the extrapolation of past values. Learn more

Google Embed this chart

Pros of the approach

- Simple to use
- Free
- Time lag only 2-3 days
- Local areas highlighted
- SOME time series do appear to follow the Official Statistics on nationals (Polish, Lithuanian, Romanian) quite closely

Cons of the approach

- Minimal methodology and metadata available – seriously limits our ability to understand and validate what we are seeing
- Use of sample restricts its use at small geographies – and quality of local data is limited by geographical precision issues (e.g. IP address location)
- Frequent changes to methodology and tools
- Discontinuities?
- Search indices instead of empirical volumes
- Lack of consistency with official definitions
- Observed results may be caused by external factors that are irrelevant to topic of interest!

Other thoughts on applications

- There may be some application in identifying potential for sudden changes in migration – before they actually happen!
- How this might be translated into actual projections of numbers is far from clear at this stage

http://www.google.com/insights/search/#q=london&geo=NP&cm

Google Insights for Search - ...

Google Insights for Search beta

+1 Help | Sign in | Download as CSV | English (UK)

Compare by	Search terms	Filter
Search terms Locations Time Ranges	Tip: Use quotation marks to match an exact phrase ("table tennis"). london + Add search term	Web Search Nepal / All sub-regions 2004 - present All Categories Search

Web Search Interest: london

Nepal, 2004 - present

The categorisation taxonomy of Google Insights for Search has been updated during December 2011. Learn more

An improvement to our geographical assignment was applied retroactively from 1/1/2011. Learn more

Interest over time

forecast | News headlines

How can I see numbers?

The last value on the graph is based on partial data and may change. Learn more

Google Embed this chart

Regional interest

Town/City

http://www.google.com/insights/search/#q=londres&geo=ES&cm

Google Insights for Search - ...

Google Insights for Search beta

+1 Help | Sign in | Download as CSV | English (UK)

Compare by

- Search terms
- Locations
- Time Ranges

Search terms

Tip: Use quotation marks to match an exact phrase ("table tennis").

londres

+ Add search term

Filter

Web Search

Spain | All sub-regions

2004 - present

All Categories

Search

Web Search Interest: londres

Spain, 2004 - present

Categories: Travel (25-50%), News (0-10%), Business & Industrial (0-10%) , more...

The categorisation taxonomy of Google Insights for Search has been updated during December 2011. Learn more

An improvement to our geographical assignment was applied retroactively from 1/1/2011. Learn more

Interest over time

forecast | News headlines

How can I see numbers?

The last value prior to the forecast is based on partial data and may change. Learn more

Future values are based exclusively on the extrapolation of past values. Learn more

Google Embed this chart

Initial thoughts on findings

- Promising early research - but limited by the lack of information about Insights methods and data quality
- May have some potential use as part of QA for dispersal of populations across the UK / sudden changes to migration patterns
- Could be enhanced by more information about language in which search terms were submitted, rather than using simple words or phrases
- Lots more work needed!