UKCRAWL: DATA OVERVIEW

Introduction

UKCrawl provides a complete collection of top-level domain (TLD) URLs (e.g. bbc.co.uk) that use the '.uk' country code top-level domain (ccTLD), geolocated using their most frequently occurring postcode coordinate location.

Data Format

This data comes split by year (e.g. `ukcrawl-2019.parquet`), each file therefore contains all UK TLDs and associated postcodes that were scraped by the Common Crawl for that particular year. For any URL that was scraped multiple times in a single year, the most recent is used to retrieve postcodes. All webpages associated with these URLs are also processed in this manner, and used to associate postcodes with their respective TLD.

Data Observations

Inconsistent Crawl Sizes

When retrieving the Common Crawl data, there was a notable drop in the total number of UK URLs that were retrieved for the final archive of 2022 and throughout 2023.

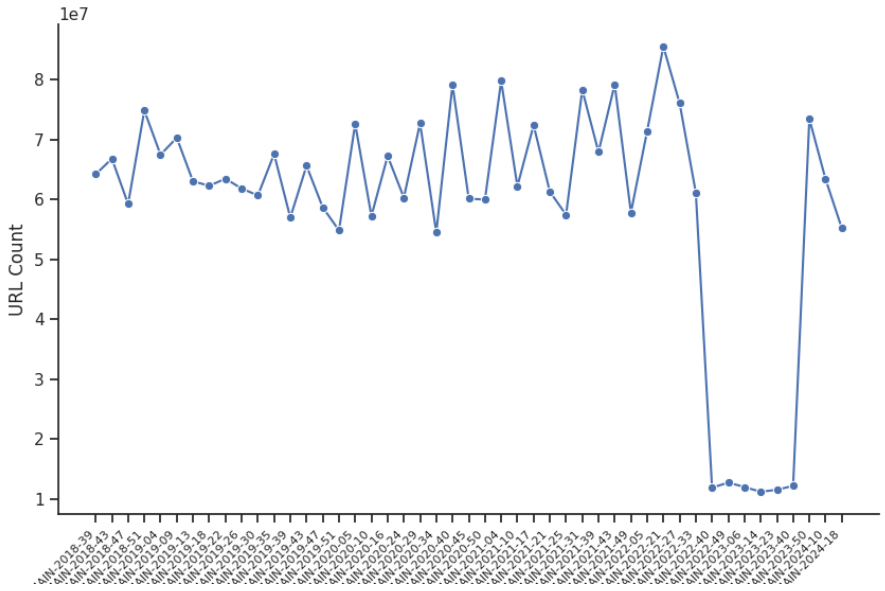

Figure 1: Number of URLs per archive

As expected, this drop in URLs also resulted in a decrease in the number of postcodes that were retrieved for these years.

	2019	2020	2021	2022	2023	2024
URLs	989,514	885,598	895,663	**768,846**	**428,967**	554,723

The Common Crawl team have confirmed that an issue with their crawl configuration during these archives resulted in the exclusion of both the 'co.uk' and 'org.uk' 2-level ccTLDs. The erratum note may be found here.

Number of Postcodes per Website

The number of postcodes found per website is highly skewed; many websites have only one postcode associated with them (11.86%), while certain websites have a disproportionally large number of postcodes. For example, the website 'floralparadise.co.uk' has 20,648,093 associated postcodes. In total, 166,477 websites have over 1,000 associated postcodes.

The following figure shows the distribution of the number of postcodes by website for each year, excluding outliers.

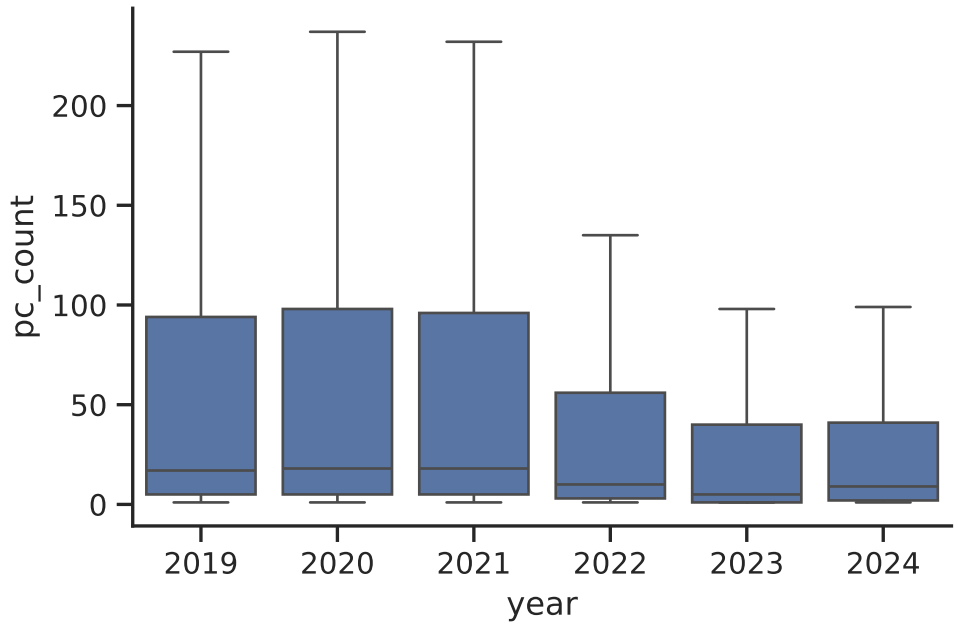

Figure 2: Number of postcodes per website

Number of Websites per Area Population

To understand the distribution of websites across the UK, we aggregate the number of URLs into Local Authority Districts (LADs) on the following figure.

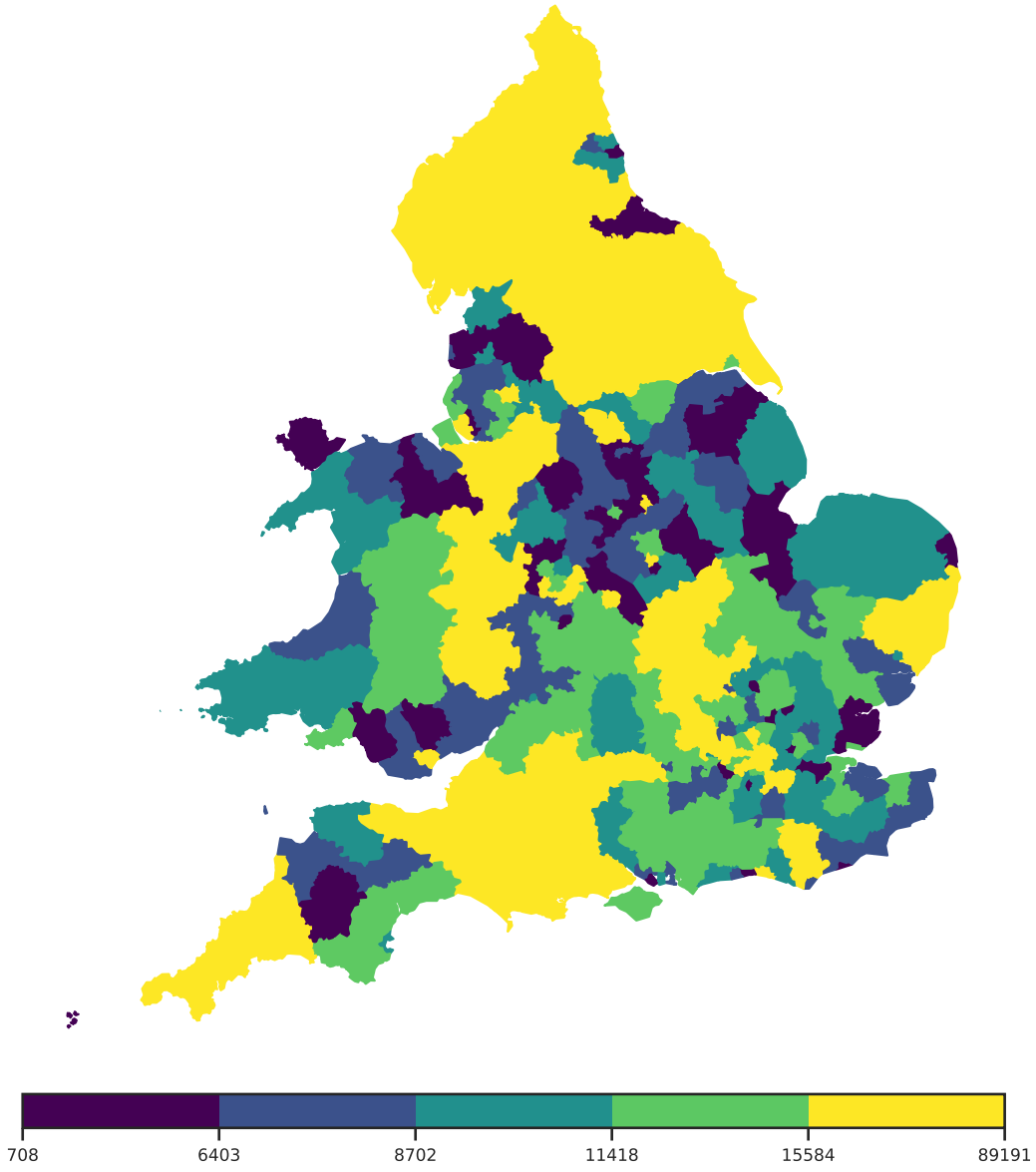

Figure 3: Number of websites per LAD