Review

An overview of research and development themes in the measurement and occurrences of polyaromatic hydrocarbons in dusts and particulates

Mark R. Cave[a,*], Joanna Wragg[a], Darren J. Beriro[a], Chistopher Vane[a], Russell Thomas[b], Matthew Riding[b], Christopher Taylor[c]

[a] British Geological Survey, Keyworth, Nottingham, NG12 5GG, UK
[b] WSP Kings Orchard, 1 Queen Street, Bristol, BS2 0HQ, UK
[c] National Grid Property Holdings Ltd, National Grid House, Warwick Technology Park, Gallows Hill, Warwick, CV34 6DA, UK

A B S T R A C T

Polycyclic aromatic hydrocarbons (PAHs) are a group of organic compounds consisting of two or more fused aromatic rings and are probably one of the most studied groups of organic chemicals in environmental research. PAHs originate mainly from anthropogenic processes, particularly from incomplete combustion of organic fuels. PAHs are distributed widely in particulate matter. Due to widespread sources and persistent characteristics, PAHs disperse through atmospheric transport and exist almost everywhere. Human beings are exposed to PAH mixtures in gaseous or particulate phases in ambient air. Long-term exposure to high concentrations of PAHs is associated with adverse health problems. This review identifies the main research and development themes in the measurement and occurrences of PAHs in dusts and particulates using a new approach to carrying out a literature review where many peer-review publications have been produced. The review extracts the most important research themes from a literature search using a combination of text mining and a more detailed review of selected papers from within the identified themes.

1. Introduction

The specific interest of our research group is the measurement of human exposure to polyaromatic hydrocarbons (PAHs) derived from airborne particulates and dust derived from coal-tar contaminated soil or related sources which can subsequently be used as part of a human health risk assessment. Whilst this is a relatively well defined research area, the aim of this review was not to narrow the search for published studies to this topic but to get as wide an overview of all the literature covering all aspects of the sources, sampling methods and measurement of PAHs in dusts and particulate matter to ensure that we could make use of any of the wider research developments in our studies.

PAHs are probably one of the most studied groups of organic chemicals in environmental research. A general search of papers on the Thompson Reuters WEB OF SCIENCE™ search engine containing PAH in the title or the abstract returns over 42,000 papers published since 1970. Plotting the number of papers published per day (Fig. 1) shows that the number of publication takes a sharp increase in 1991 to about 1 paper per day with an approximate linear increase 8 papers per day in 2016.

The Thompson Reuters research categories (Fig. 2) show that over 30% of the papers come from environmental science and ecology research and 20% on chemistry research topics reflecting the fact that their effects on the environment are the main concern and that development of analytical chemistry protocols and sampling methods playa a crucial role in quantifying PAHs in environmental matrices.

This very large data set which contains a huge store of information but accessing specific topics within this data has two main drawbacks:

i) It is very difficult to get an overview of the trends in a data set of thousands of papers as it is impossible to read and digest the information in that many abstracts let alone read all the papers; and
ii) Even if a team of reviewers were able to read the papers and produce an overview review within a period of a few weeks the review would be out of date as new papers are published daily (Fig. 1).

One obvious way of making the task more manageable is to refine the search to be more specific for a given research area. In this study we are interested in PAH in airborne particulate matter therefore, in addition to PAH we searched specifically for dusts or particulates or PM10 or PM2.5 or geogenic dusts but excluded cosmic or interstellar or galactic or diesel or sediments or leaf or leaves or leafy (for the full search
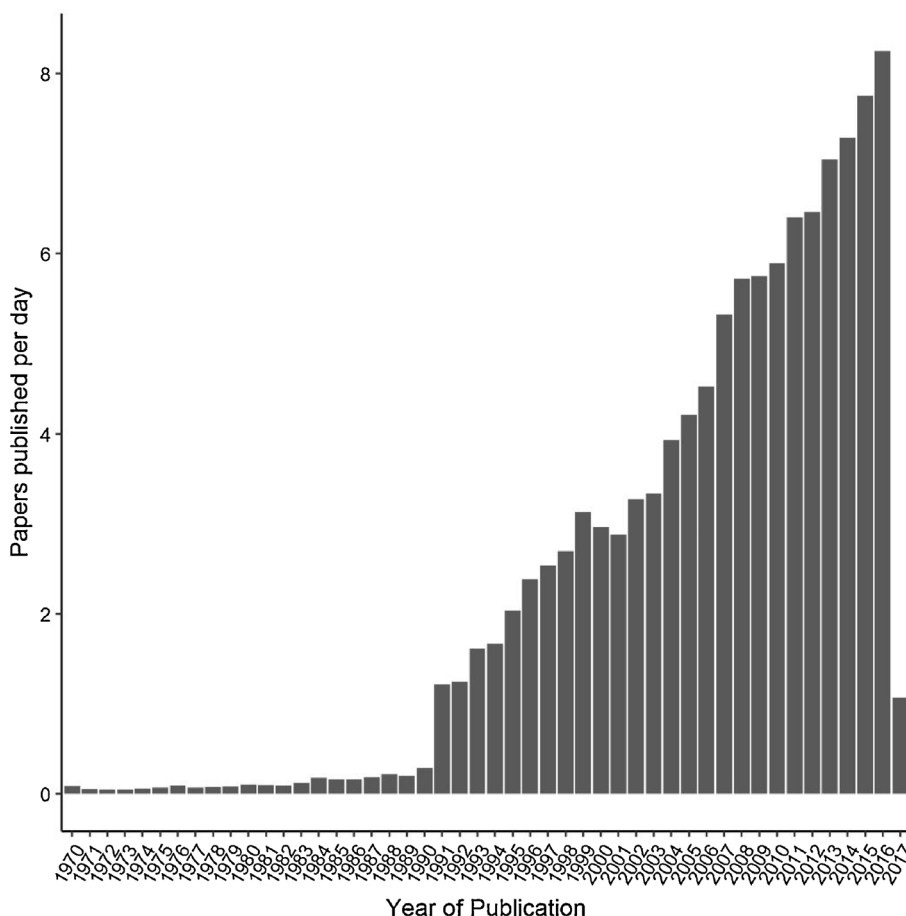
Fig. 1. Number of papers published per day with PAH in the title or abstract.

specification see the supplementary material). This reduced the number of papers down to 3041. Whilst inhalation of PAH contaminated soil particulates entrained in air at contaminated land sites is the particular focus of this study further refining the literature would exclude a large amount of material which may not be exactly aligned to our study but may contain important information that can have direct application to our research. It is for this reason that we decided to try and review this larger literature body.

Fig. 3 shows the number of papers published per year for the refined PAH search which broadly follows the same pattern as the unrefined search with a linear increase in papers published between 1991 and 2016 but the rate of publication is between ca.0.7 and 5 papers per week. Despite a reduction in papers by a factor of 15, more than 3000 papers is still a large number of papers to read and be reviewed individually.

Text mining is the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. The biological sciences, in particular, are making use of this method of extracting data from large amounts of scientific literature [1].

The aim of this study is to use a combination of data and text mining to review the relevant key scientific trends in the large volume of PAH particulate matter literature in a rapid and efficient manner and to drill down to give a more detailed review of the key papers covering PAH content of particulates.

2. Methods

The study will use the information supplied with the Thomson Reuters saved search as outlined in Table S1 and saved as a tab

delimited text file.

The tab delimited text file will be split into the Thompson Reuters defined research areas so that differences between the research viewpoints and the overall data can be examined.

The review process will then proceed as follows:

Review the timescales, journal names and number of citations of the papers in each group;

Carry out text mining of the paper titles to find commonly occurring word pairs using the "tidytext" package in the R programming language [2] and visualising these as network graphs us the R "igraph" package [3] to show the underlying trends within the different research areas;

From each of the research areas select 5 of the highest cited papers most strongly associated with measuring PAH in particulate matter and carry out a detailed review by reading the original text of each paper; and

Use the R programming language [4] text string search packages "stringr" [5] and "stringi" [6] to find and summarise occurrences of key phrases and associated numerical entries identified as being important by the text mining and the specific detailed reviews in the abstracts of all of the papers obtained from the Thompson Reuters review search.

3. Results

3.1. All research areas

Using the addresses of the corresponding authors to identify the countries associated with each paper, Fig. 4 shows the top twenty countries with USA, China and Italy being the main contributors but there is good coverage from most of the major continents apart from Australasia. Fig. 5 shows the number of publications categorised by