Regulatory Toxicology and Pharmacology 99 (2018) 251–259

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

Modeling of cigarette smoke constituents - From intense to less intense smoking regime

Thomas Verron*, Rémi Julien, Xavier Cahours, Stéphane Colard

SEITA, Imperial Brands, 143 Boulevard Romain Rolland, 75685, Paris, France

ARTICLE INFO

Keywords:
Multiple linear regressions
Computational process
Mainstream cigarette smoke
Smoking regimes

ABSTRACT

Since it was first required to measure and to report NFDPM and nicotine yields in a limited number of countries, there has been an increasing trend for more testing and reporting requirements. Historically, the ISO 3308 smoking regime has been used to determine NFDPM and nicotine yields. However recommendations from the World Health Organization, now include the use of two smoking regimes such as the ISO 3308 and the WHO TobLabNet Official Method SOP01, the latter being considered as an intense smoking regime. Considering the increase in data produced and similarities between some smoke constituents formed during combustion, we explored possible correlations between emissions under intense and less intense smoking conditions. A set of 22 commercial cigarettes was tested. Eighty five smoke constituents were determined under both intense and less intense regimes. In addition 36 tobacco constituents, 14 cigarette design parameters and eight cigarette burning features were determined. A computational process was designed to implement multiple linear regression analyses enabling the identification of the best subsets of explanatory variables among emissions under intense conditions, cigarette design parameters, tobacco constituents and burning parameters. We succeeded in building simple linear models, involving four to six variables, while reaching satisfactory goodness of fit and R-squared values ranging from 0.87 to 1.00. Our findings suggest, in the range of products tested, that the additional data gained by using a second smoking regime does not necessarily increase the volume of information and consequently does not necessarily improve knowledge. This study supports the premise that the application of two smoking regimes does not produce a more comprehensive product characterisation compared to using one.

1. Introduction

Several regulatory and scientific advisory bodies have required cigarette smoke constituent yield data for both reporting and product comparison purposes. The European Union (EU) and many other worldwide governmental authorities have introduced regulations on smoke constituents that, for example, require manufacturers to report Nicotine Free Dry Particulate Matter (NFDPM) or 'tar' (T), nicotine (N) and carbon monoxide (CO) yields or set yield limits which must not be exceeded (European Union Directive, 2001). TNCO yields are measured on the collected smoke as prescribed by ISO testing methods (ISO, 1999b; ISO, 2000a; ISO, 2000b; ISO, 2000c; ISO, 2009b), which are intended to provide a means of ranking cigarettes in terms of TNCO yields under fixed puffing conditions. The tobacco industry has reported cigarette TNCO yield data for many years. However, Parties signatory to the World Health Organization Framework Convention on Tobacco Control (WHO FCTC) and the US Food and Drug Administration (FDA), are currently proposing to increase regulation of tobacco and cigarette smoke constituents (Food and Drug Administration, 2016; World Health Organization, 2008). The FDA has published a list of 93 harmful and potentially harmful constituents (HPHCs) in tobacco products and tobacco smoke (Food and Drug Administration, 2012b), and published draft guidance on the reporting of an abbreviated list of HPHCs for which analytical protocols are assumed to be well established and widely available (Food and Drug Administration, 2012a). In 2015, the WHO Study Group on Tobacco Product Regulation (TobReg) established a non-exhaustive priority list of 39 cigarette constituents and emissions (World Health Organization, 2015). FDA recommends that the yield of each HPHC in cigarette smoke should be determined by both less intense and intense smoking conditions. Historical conditions are described in the ISO 3308 testing method, that require a 35 mL puff volume, a 2s puff duration and a 60s puff interval with no blocking of filter ventilation holes. Intense conditions are described in the WHO TobLabNet Official Method SOP01 (WHO Tobacco Laboratory Network, 2012) testing method, that requires a 55 mL puff volume, a 2s puff duration and a 30s puff interval with the blocking of filter ventilation

* Corresponding author. SEITA, Imperial Brands, 143 Boulevard Romain Rolland, 75685, Paris, France.
E-mail address: thomas.verron@fr.imptob.com (T. Verron).

https://doi.org/10.1016/j.yrtph.2018.09.015
Received 15 January 2018; Received in revised form 10 September 2018; Accepted 12 September 2018
Available online 15 September 2018
0273-2300/

holes. Both testing methods require a bell-shaped puff profile. TobReg and WHO Tobacco Laboratory Network (TobLabNet) suggest that more than one machine regime is required to characterize smoke constituents adequately (World Health Organization, 2008). The two smoking regimes are expected to provide information about the range of deliveries of HPHCs possible for each cigarette product (Food and Drug Administration, 2012b).

It is unknown which regime would provide the best characterisation for regulation as none of the smoking regimes represent human behaviour. None are likely to produce data that will be markedly associated with human exposure or risk, either for individual smokers or for population-level differences between products. In fact, the purpose of the testing regimes is mainly to characterize how products perform under a specific set of smoking conditions. The question then turns to how many smoking regimes are required to obtain accurate product characterisation. Few publications show potential correlations between smoking regimes. Purkis et al. (2013) compared TNCO yields from cigarettes obtained under 16 different smoking regimes. They showed that TNCO yields formed part of a continuous function linked with puffing intensity (the product of puff volume and puff frequency) and total puff volume (the product of puff volume and puff number). As such, the characterisation of cigarette products can be achieved using TNCO yields generated with a single smoking regime. Colard et al. (2014) showed that TNCO yields determined under different smoking regimes, with and without ventilation blocking, are linearly related to the difference between the smouldering time (cigarette combustion with no puffing) and the smoking time (cigarette combustion with puffing). The authors concluded that the obtained correlations provide an approach to predict TNCO yields from one smoking regime to another. Few publications measured smoking regimes correlations for smoke constituents other than TNCO. King et al. (2007) published a multidimensional analysis of 13 selected mainstream smoke constituents generated under both ISO 3308 and Health Canada Intense (Health Canada, 2000) machine smoking regimes using 15 Australian and 21 Canadian cigarette products. Correlations among the smoke constituents were computed and stepwise regressions used to predict emissions under Canadian Intense machine-smoking conditions from the ISO 3308-generated smoke data. Using linear regression, they reported that NFDPM, nicotine, and CO yields under ISO 3308 conditions and filtration efficiency were good predictors of intensive condition adjusted emissions of vapour phase constituents but not for particulate phase constituents. Colard (2015) proposed a novel testing scheme involving the determination of number of puffs and smoking times under two different smoking regimes and inputting this data into a cigarette burning model. This enabled the author to characterize the burning process and provided an extensive set of information such as the mean smoulder rate between puffs and the mass of tobacco burnt during puffs regardless of the smoking regime applied. Good correlations were observed between the mass of tobacco burnt during puffs and the yields of 18 HPHCs in cigarette smoke (Food and Drug Administration, 2012b), suggesting that yields determined from one regime are sufficient to establish the relationships between yields and smoking intensity. Pazo et al. (2016) analyzed 21 volatile organic compounds (VOCs) in mainstream cigarette smoke of 50 commercial cigarettes and two research cigarettes. They reported that, for both smoking regimes under intense and less intense conditions, mainstream smoke VOC amounts among the different products were strongly correlated between the majority of the analytes. However, mono-aromatic compounds were found to increase disproportionately compared to unsaturated, nitro, and carbonyl compounds under the intense smoking regime. The authors suggested that a possible cause for the disproportionate increase in monoaromatic compounds could be increased pyrolysis under low oxygen conditions associated with the intense smoking conditions.

Our current study considered a large set of variables as a set of candidate predictors for modeling of smoke constituents under ISO 3308 conditions. The pool constituted four main blocks of information: smoke constituents under intense conditions (Health Canada, 1999), tobacco filler constituents, cigarette design parameters and cigarette burning parameters.

A computational process composed of two steps, selection and modeling, has been implemented. The selection step aimed to find an appropriate subset of predictors, from the four blocks of variables. The second step aimed to model smoke constituents under ISO 3308 conditions with multiple linear regressions. The interpretation of these results enables us to assess the added value of the requirement of using two smoking regimes.

2. Materials and methods

2.1. Cigarettes

A selection of 22 commercial cigarettes, manufactured in 2012 for the US market, were tested. All products were blended with an American style blend of burley, flue cured and oriental tobacco and some products had a blend with menthol flavors. A range of cigarette formats were assessed including cigarettes of different lengths (King Size, 100 mm, 120 mm), diameters (standard vs slim) and filter ventilations. Table 1 summarizes the design characteristics for the cigarettes used.

2.2. Mainstream smoke testing

Testing was performed by Arista Laboratories Inc. (Richmond, Virginia), a commercial laboratory contracted to Imperial Tobacco Limited, between October 2012 to April 2013. Cigarette samples were conditioned according to ISO 3402 (ISO, 1999a), and both ISO 3308 and HC T-115 (Health Canada, 1999) smoking conditions (equivalent to the WHO TobLabNet Official Method SOP01) were applied to generate mainstream smoke. Twenty replicates were performed for NFDPM, nicotine and carbon monoxide, and seven replicates were performed for all other mainstream smoke constituents. A total of 85 smoke constituents were analyzed using ISO 17025-accredited methods. Table 2 provides a list of analyzed constituents with their associated analytical method or standard. In order to avoid Cambridge filter pad saturation for the intense smoking conditions, an appropriate adjustment was made. Five cigarettes per replicate were used for the ISO 3308 smoking conditions and from two to three cigarettes per replicate were used for the intense smoking regime, accordingly to the product yields.

The 85 mainstream smoke constituents under intense conditions were used as the first block of information to build the multiple linear regressions.

Table 1
Cigarette design characteristics.

Characteristic	Range
ISO machine-method pack printed NFDPM yield	5–13 mg/cig
ISO machine-method pack printed nicotine yield	0.5–1.0 mg/cig
Filter construction	Filtered brands
Filter types	Acetate filter material
Filter ventilation	0–56%
Cigarette length	82–118 mm
Cigarette diameter	6.1–7.8 mm
LIP paper	All LIP
Cigarette paper permeability[a]	60-82 CORESTA units
Menthol flavoring	9 mentholated/13 non-mentholated
Tobacco weight	0.42–0.76 g/cig
Tobacco blend	American blend

[a] Cigarette paper permeability measured between low permeability band regions.

Table 2
List of 85 smoke constituents generated under intense and less intense conditions, and associated method of analysis.

Smoke constituents (#85)	Test procedure[a]	Separation & detection methods	Extraction solvent & solution
Carbonyls (#7)			
Acetaldehyde, Acetone, Acrolein, Crotonaldehyde, Formaldehyde, Methyl ethyl ketone, Propionaldehyde	HC T-104	UHPLC[b] & UVD[c]	Acetonitrile/Acidified, 25 mM 2,4 dinitrophenylhydrazine
Heterocyclic aromatic amines (#8)			
A-α-C (2-amino-9H-pyrido ([2,3-b]indole) Glu-P-1 (2-amino-6-methyldipyridol[1,2-a:3′,2′-d]imidazole) Glu-P-2 (2-aminodipyridol[1,2-a:3′,2′-d]imidazole) IQ (2-amino-d-methylimidazo[4,5-f]quinolone) MeA-α-C (2-amino-3-methyl-9H-pyrido[2,3-b]indole) PhIP (2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine) Trp-P-1 (3-amino1,4-dimethyl-5H-pyrido[4,3-b]indole) Trp-P-2 (1-methyl-3-amino-5H-pyrido[4,3-b]indole)	NA	UHPLC[b] & MS[d] with ESI[e]	Water & 0.1% formic acid
Metals (#8)			
Arsenic, Beryllium, Cadmium, Chromium, Cobalt, Lead, Nickel, Selenium	HC T-109	ICPMS[f]	Methanol & Nitric acid, hydrogen peroxide
Nitrosamines (#8)			
NDEA (N-Nitrosodiethylamine), NDMA (N-Nitrosodimethylamine) NMEA (N-Nitrosomethylethylamine) NMOR (N-Nitrosomorpholine), NPIP (N-Nitrosopiperidine) NPYR (N-Nitrosopyrrolidine)	NA	GC[g]-MS[d] & MS[d]	Dichloromethane & Water
NNK (4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone) NNN (N-Nitrosonornicotine)	NA	HPLC[h] & MSMS[i]	Type I or HPLC[h] grade water & 100 mM aqueous ammonium acetate
Others (#9)			
Ammonia	HC T-101	HPLC[h] & Non-suppressed conductivity	Water & 10 mM methanesulforic acid
Caffeic acid	NA	LC[j] & MS[d] with ESI[e]	Methanol
Carbon monoxide	ISO 8454	NDIR[k]	NA
Ethyl carbamate	NA	GC[g] & MS[d]	Ethyl acetate
Hydrazine	NA	UHPLC[b] & UVD[c]	Acetonitrile & 0.1M benzaldehyde
Hydrogen cyanide	HC T-107	CFA[l] & Photometric	Water & 0.1M aqueous sodium hydroxide
NFDPM (Nicotine Free Dry Particulate Matter), TPM (Total Particulate Matter)	ISO 4387	NA	Isopropanol & Isopropanol with heptadecane and ethanol
Nicotine	ISO 10315	GC[g] & FID[m]	Isopropanol & Isopropanol with heptadecane and ethanol
Phenolic compounds (#4)			
Catechol, Cresols (m + p-cresol), Cresols (o-cresol), Phenol	HC T-114	HPLC[h] & Fluorescence	Water & 1% v/v acetic acid (with 2.5% methanol)
Polycyclic aromatic amines (#6)			
1-Aminonaphthalene, 2-Aminonaphthalene, 4-Aminobiphenyl, 2,6-Dimethylaniline, o-Anisidine, o-Toluidine	HC T-102	GC[g] & MS[d]	Methylene chloride & 1.6N hydrochloric acid
Polycyclic aromatic hydrocarbons (#15)			
5-Methylchrysene, Benz[a]anthracene, Benz[j]aceanthrylene, Benzo[a]pyrene, Benzo[b]fluoranthene, Benzo[c]phenanthrene, Benzo[k]fluoranthene, Chrysene, Cyclopenta[c,d]pyrene, Dibenzo[a,e]pyrene, Dibenzo[a,h] anthracene, Dibenzo[a,h]pyrene, Dibenzo[a,i]pyrene, Dibenzo[a,l]pyrene, Indeno[1,2,3-cd]pyrene	NA	GC[g] & MS[d]	Cyclohexane & Methanol
Selected volatiles (#11)			
2-Nitropropane, Acrylonitrile, Benzene, Benzo[b]furan, Ethylbenzene, Furan, Naphthalene, Nitrobenzene, Nitromethane, Styrene, Toluene	NA	GC[g] & MS[d]	Methanol
Semivolatiles (#3)			
Acetamide, Acrylamide	NA	GC[g] & MS[d]	Methanol
Quinoline	HC T-112	GC[g] & MS[d]	Methanol & Triethylamine
Volatiles (#6)			
1,3-Butadiene, Ethylene oxide, Isoprene, Propylene oxide, Vinyl acetate, Vinyl chloride	NA	GC[g] & MS[d]	Methanol

[a] Reference is made to Health Canada official method (HC) or ISO standard method.
[b] UHPLC, ultra-high performance liquid chromatography.
[c] UVD, ultraviolet detector.
[d] MS, mass spectrometry.
[e] ESI, electrospray ionisation.
[f] ICPMS, inductively coupled plasma mass spectrometry.
[g] GC, gas chromatography.
[h] HPLC, high performance liquid chromatography.
[i] MSMS, mass spectrometry-mass spectrometry.
[j] LC, liquid chromatography.
[k] NDIR, nondispersive infrared.
[l] CFA, continuous flow analyzer.
[m] FID, flame ionisation detector.

Table 3
List of 36 tobacco filler constituents representing the second block of information used as predictor candidates.

Tobacco filler constituents (#36)	Test procedure[a]	Separation & detection methods	Extraction solvent & solution
Alkaloids (#3)			
Anabasine, Nicotine, Nornicotine	NA	GC & MS	Methanol & Methanol with quinoline
Carbonyls (#3)			
Acetaldehyde, Crotonaldehyde, Formaldehyde	NA	UHPLC & UVD	Phosporic acid & Acidified, 25 mM 2,4 dinitrophenylhydrazine
Metals (#8)			
Arsenic, Beryllium, Cadmium, Chromium, Cobalt, Lead, Nickel, Selenium	HC T-306	ICPMS	Nitric acid
Nitrosamines (#9)			
NDELA	NA	GC & NCD[b]	Ethyl acetate & Water/Ethyl acetate
NDMA, NMEA, NMOR, NPIP, NPYR	NA	GC-MS & MS	Dichloromethane
NNK, NNN	NA	HPLC/UHPLC & MSMS	Type I or HPLC grade water & 100 mM aqueous ammonium acetate
NSAR	NA	GC & NCD[b]	Ethyl acetate & 18N sulfuric acid, 1% ammonium sulfamate
Others (#5)			
Aflatoxin B1	NA	HPLC/UHPLC & MSMS	Methanol/Water
Ammonia	HC T-302	HPLC & Non-suppressed conductivity	Water & 10 Mm Methanesulforic acid
Coumarin	NA	UHPLC & MSMS	Methanol
Ethyl carbamate	NA	GC & MS	Ethyl acetate & Water
Mercury	HC T-306	Inductively coupled plasma & MS	Nitric acid
Polycyclic aromatic hydrocarbons (#8)			
Benz[a]anthracene, Benzo[a]pyrene, Benzo[b]fluoranthene, Benzo[k] fluoranthene, Chrysene, Dibenzo[a,h]anthracene, Indeno[1,2,3-cd] pyrene, Naphthalene	NA	GC & MS	Methanol & Cyclohexane

[a] Reference is made to Health Canada official method (HC) or ISO standard method.
[b] NCD, nitrogen chemiluminescence detector.

2.3. Tobacco filler testing

The tobacco filler constituents were analyzed by Arista Laboratories Inc. (Richmond, Virginia), between October 2012 to April 2013. Although the ISO 5725-6 standard (ISO, 2001) recommends the execution of two to four replicates for validated analytical methods, the contracted laboratory performed seven replicates for such analysis. The list of the 36 tobacco filler constituents analyzed in this study is given in Table 3 with the corresponding method or standard. This second list constituted the second block of information for building the multiple linear regressions.

2.4. Cigarette design

Fourteen parameters related to the cigarette designs were included. Six parameters were related to the dimensional description: diameter (ISO, 2013), filter length, tipping length, butt length, tobacco length and total length. Additionally, seven other parameters describing tobacco and non-tobacco materials were used: tobacco weight, tobacco dry weight, tobacco density, filter and cigarette paper ventilations (ISO, 2003), cigarette paper grammage and cigarette paper permeability (ISO, 2009a). To complete the cigarette description, smoldering rate was determined from measuring the rate of tobacco rod weight loss during combustion, and the conversion of the measured weight into the corresponding length based on an internal standard operating procedure.

The set of 14 cigarette design parameters formed the third block of information for building the multiple linear regressions.

2.5. Cigarette burning parameters

Cigarette smoking consists of successive steps of active burning during each puff then smouldering between each puff. A cigarette burning model (Colard, 2015; Colard et al., 2014) was developed to describe and reproduce this sequential process taking into account some key parameters such as smoldering rate, filter and paper ventilation, puff duration, puff interval, puff volume, tipping length, cigarette length and butt length. The following output parameters can then be readily deduced: puff number, length of rod actively burnt during all puffs, weight of tobacco actively burnt (by puffing), mean mass per puff and smoking time, independent of smoking regime applied.

Finally, a set of eight parameters related to the cigarette burning was selected to form the fourth block of information and used as explanatory variables for building the multiple linear regressions.

2.6. Statistical approach

Four blocks of information were then considered as predictor candidates for attempting to establish statistical relationships with smoke constituents under ISO 3308 conditions. The set of explanatory variables was composed of smoke constituents analyzed under intense conditions (Block n°1, 85 variables), of cigarette tobacco filler constituents (Block n°2, 36 variables), of cigarette design parameters (Block n°3, 14 variables) and of burning parameters (Block n°4, 8 variables).

A computational process was developed to identify the best relationships between smoke constituents under ISO 3308 conditions and the explanatory variables. Overall, a total of 143 explanatory variables (x_i) was available to build relationships (Eq. (1)) with each of the smoke constituents (Y).

$$Y = \beta_0 + \sum_{i=1}^{I} \beta_i x_i \quad (1)$$

with $I \in \{1, ..., 6\}$ *and* $x_i \in \{Block1 \cup Block2 \cup Block3 \cup Block4\}$

The computational process, was applied iteratively for each smoke constituent under ISO 3308 smoking conditions. In a first step, a comprehensive search of the best subsets of up to six explanatory variables was implemented for predicting the smoke constituents. All tested combinations from all sizes up to six variables among Y, required more than 10 billion multiple regression models for each smoke constituents. The process was voluntarily limited to a maximum of six

Table 4
Contingency table of brands smoking emissions below the limit of quantification (LOQ).

Smoke constituents (#40)	LOQ			Number of brands below LOQ		
	ISO 3308 (ng/cig)	ISO 20778 (ng/cig)	Tobacco (ng/g)	ISO 3308	ISO 20778	Tobacco
Heterocyclic aromatic amines (#6)						
Glu-P-1	8	8	NA[a]	22	22	–
Glu-P-2	12	12	NA[a]	22	22	–
IQ	12	12	NA[a]	22	22	–
PhIP	24	24	NA[a]	22	22	–
Trp-P-1	12	12	NA[a]	22	22	–
Trp-P-2	4	4	NA[a]	22	22	–
Metals (#6)						
Arsenic	2.75	2.75	NA[a]	6	2	–
Beryllium	0.5	0.5	198	22	22	22
Chromium	4.25	4.25	NA[a]	22	22	–
Cobalt	0.5	0.5	NA[a]	22	22	–
Nickel	4.75	4.75	NA[a]	22	22	–
Selenium	2.25	2.25	200	21	20	22
Nitrosamines (#8)						
NDEA	3.2	8	NA[a]	22	22	–
NDELA	NA[a]	NA[a]	167	–	–	22
NDMA	3.2	8	4	3	0	22
NMEA	3.2	8	4	12	2	22
NMOR	16	40	10	22	22	22
NPIP	3.2	8	8.8	22	22	22
NPYR	NA[a]	NA[a]	10	–	–	22
NSAR	NA[a]	NA[a]	25	–	–	22
Others (#6)						
Aflatoxin B1	NA[a]	NA[a]	11.2	–	–	22
Ammonia	$5.16.10^3$	$5.16.10^3$	NA[a]	3	0	–
Caffeic acid	35.10^3	35.10^3	NA[a]	4	0	–
Coumarin	NA[a]	NA[a]	500	–	–	22
Ethyl carbamate	2.5	2.5	50	22	22	22
Hydrazine	15	15	NA[a]	22	21	–
Polycyclic aromatic hydrocarbons (#10)						
5-Methylchrysene	$1.6.10^3$	$2.67.10^3$	NA[a]	22	21	–
Benz[j]aceanthrylene	4.10^3	$6.67.10^3$	NA[a]	8	13	–
Benzo[c]phenanthrene	$1.6.10^3$	$2.67.10^3$	NA[a]	17	14	–
Benzo[k]fluoranthene	$1.6.10^3$	$2.67.10^3$	10	2	0	17
Dibenzo[a,e]pyrene	$1.6.10^3$	$2.67.10^3$	NA[a]	22	22	–
Dibenzo[a,h]anthracene	$1.6.10^3$	$2.67.10^3$	10	22	21	22
Dibenzo[a,h]pyrene	$1.6.10^3$	$2.67.10^3$	NA[a]	22	22	–
Dibenzo[a,i]pyrene	4.10^3	$6.67.10^3$	NA[a]	22	22	–
Dibenzo[a,l]pyrene	$1.6.10^3$	$2.67.10^3$	NA[a]	22	22	–
Indeno[1,2,3-cd]pyrene	NA[a]	NA[a]	10	–	–	16
Selected volatiles (#2)						
2-Nitropropane	95	238	NA[a]	22	22	–
Nitrobenzene	NA[a]	100	NA[a]	22	22	–
Volatiles (#2)						
Vinyl acetate	200	NA[a]	NA[a]	1	–	–
Vinyl chloride	100	250	NA[a]	22	22	–

[a] Not Available information.

explanatory variables in order to avoid overfitting issues (Draper and Smith, 1981; Frost, 2015; Hawkins, 2004) and to have reasonable computational time. Indeed, the complexity of the model and the number of possible subsets increase with the number of predictors.

In a second step, a branch-and-bound algorithm (Akaike, 1998; Schwarz, 1978) was applied in a computational time reduction objective. This algorithm allowed the reduction of the amount of computation involved in examining subset and the finding of the best subsets without examining all those available. A Bayes Information Criterion (BIC) (Furnival and Wilson, 2000; Hand, 1981) was implemented as well in order to choose the appropriate dimensionality of the model that contain all that is necessary for the modeling. This criterion is accepted for dealing with the trade-off between the goodness of fit and the complexity of the model.

This computational process was implemented in R (R Core Team, 2016) using the package “Leaps” (Lumley, 2006), and required less than five minutes per smoke constituent.

3. Results and discussion

3.1. Levels of chemical compounds

A significant proportion of the smoke constituents had concentrations lower than the limits of quantification (LOQ) of the analytical methods. Table 4 lists for each constituent the number of products having emissions lower than the LOQs.

In order to avoid overfitting models due to low number of observations versus parameters, we have considered a minimum of 11 quantified products (half of the whole set of products) necessary for attempting to establish relevant modeling of smoke constituents under ISO 3308 conditions. Consequently, yields with more than 11 products below the LOQ were not considered in the statistical evaluation. This corresponded to 27 out of the 85 (32%) smoke constituents that were not modeled. Finally 58 smoke constituents under ISO 3308 conditions were potential candidates for modeling. The same rule was applied to smoke constituents under intense conditions, where 27 out of the 85

Table 5
Description of the best modeling of smoke constituents under ISO 3308 conditions and selected according to the Bayes Information Criteria (BIC).

Smoke constituents (#58)	Min; Max	R2	Number of explanatory variables in …				
			Total	Intense Block 1	Tobacco Block 2	Design Block 3	Burning Block 4
Carbonyls (μg/cig) (#7)							
Acetaldehyde	287.8; 891.2	0.979	5	3	–	1	1
Acetone	128.1; 406.2	0.989	6	4	–	2	–
Acrolein	26.12; 84.14	0.964	5	3	–	1	1
Crotonaldehyde	4.81; 26.87	0.990	6	4	–	1	1
Formaldehyde	3.319; 31.43	0.978	6	3	–	1	2
Methyl ethyl ketone	31.21; 106.4	0.989	6	4	–	2	–
Propionaldehyde	23.11; 73.58	0.992	6	3	–	3	–
Heterocyclic aromatic amines (ng/cig) (#2)							
A-α-C	29.82; 76.98	0.985	6	2	1	1	2
MeA-α-C	10.58; 35.03	0.988	6	4	1	1	–
Metals (ng/cig) (#3)							
Arsenic	3.017; 4.915	0.998	6	3	–	3	–
Cadmium	18.56; 69.46	0.991	6	4	1	1	–
Lead	7.1; 33.68	0.986	6	1	4	1	–
Nitrosamines (ng/cig) (#4)							
NDMA	2.92; 8.571	0.988	6	3	1	1	1
NPYR	3.579; 14.9	0.871	4	3	–	–	1
NNK	23.82; 193.9	0.967	6	3	–	3	–
NNN	41.34; 213.5	0.996	6	3	–	3	–
Others (#7)							
Ammonia *(μg/cig)*	6.123; 18.39	0.900	4	1	–	2	1
Caffeic acid *(μg/cig)*	36.51; 190.7	0.982	6	5	–	1	–
Carbon monoxide *(mg/cig)*	4.244; 13.2	0.992	6	3	1	1	1
Hydrogen cyanide *(μg/cig)*	46.89; 204	0.989	6	4	–	1	1
NFDPM *(mg/cig)*	4.678; 13.63	0.994	6	3	1	2	–
TPM *(mg/cig)*	5.154; 16.41	0.994	6	2	1	2	1
Nicotine *(mg/cig)*	0.3683; 0.9322	0.996	6	4	1	1	–
Phenolic compounds (μg/cig) (#4)							
Catechol	20.48; 57.94	0.995	6	3	–	1	2
Cresols (m + p-cresol)	3.986; 13.35	0.995	6	4	–	1	1
Cresols (o-cresol)	1.387; 5.205	0.993	6	5	–	1	–
Phenol	4.222; 20	0.995	6	5	–	1	–
Polycyclic aromatic amines (ng/cig) (#6)							
1-Aminonaphthalene	6.389; 14.95	0.989	6	5	–	1	–
2-Aminonaphthalene	3.649; 7.83	0.990	6	5	–	1	–
4-Aminobiphenyl	0.7771; 1.657	0.984	6	5	–	1	–
2,6-Dimethylaniline	1.549; 8.224	0.992	6	5	–	1	–
o-Anisidine	1.3; 3.737	0.989	6	3	–	2	1
o-Toluidine	27.09; 72.27	0.976	5	2	–	2	1
Polycyclic aromatic hydrocarbons (μg/cig) (#8)							
Benz[a]anthracene	6.226; 15.96	0.992	6	4	1	1	–
Benz[j]aceanthrylene	4.776; 17.52	1.000	6	5	–	1	–
Benzo[a]pyrene	3.757; 9.602	0.995	6	4	–	1	1
Benzo[b]fluoranthene	2.741; 7.011	0.991	6	2	1	2	1
Benzo[k]fluoranthene	1.721; 2.993	0.987	6	3	–	2	1
Chrysene	7.497; 18.54	0.993	6	4	–	2	–
Cyclopenta[c,d]pyrene	4.459; 19.67	0.959	6	6	–	–	–
Indeno[1,2,3-cd]pyrene	1.688; 4.204	0.996	6	3	1	1	1
Selected volatiles (μg/cig) (#9)							
Acrylonitrile	5.882; 17	0.987	6	4	1	1	–
Benzene	15.31; 48.84	0.989	6	3	–	2	1
Benzo[b]furan	0.05413; 0.3934	0.990	6	4	–	2	–
Ethylbenzene	2.842; 9.299	0.984	6	5	–	1	–
Furan	9.608; 28.89	0.989	6	5	–	1	–
Naphthalene	0.1398; 0.8688	0.985	6	5	–	1	–
Nitromethane	0.1086; 0.5048	0.989	6	5	–	1	–
Styrene	2.165; 9.762	0.987	6	5	–	1	–
Toluene	27.97; 80.78	0.984	6	5	–	1	–
Semivolatiles (#3)							
Acetamide *(ng/cig)*	954.9; 6614	0.992	6	3	–	2	1
Acrylamide *(ng/cig)*	777.2; 3808	0.991	6	2	1	1	2
Quinoline *(μg/cig)*	0.1177; 0.5036	0.995	6	4	1	1	–
Volatiles (μg/cig) (#5)							
1,3-Butadiene	15.9; 51.23	0.951	5	2	1	–	2
Ethylene oxide	4.924; 14.25	0.964	6	5	–	1	–
Isoprene	163.7; 565.6	0.981	6	2	1	–	3
Propylene oxide	0.4737; 3.637	0.996	6	3	1	1	1
Vinyl acetate	0.217; 0.5585	0.974	6	4	1	1	–

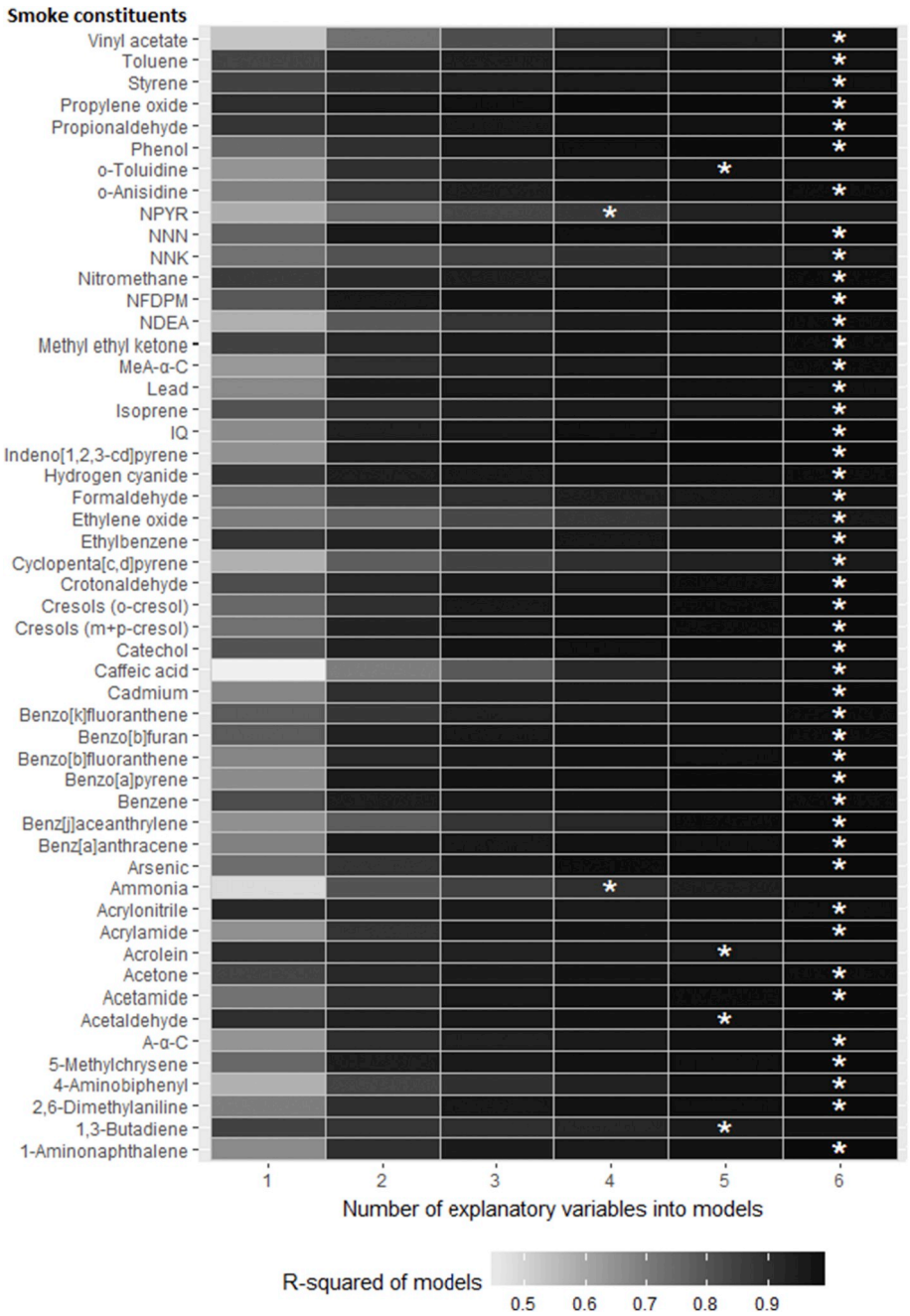

Fig. 1. Representation of the highest R-squared values. The star indicates the ultimate model selected according to the Bayes Information Criteria (BIC).

(32%) smoke constituents were not employed as explanatory variables. Applying the rule to tobacco filler resulted in 15 out of the 36 (42%) constituents being discarded from the pool of candidates for modeling of smoke constituents under ISO 3308 conditions. Overall, up to 101 explanatory variables were available and potentially usable for modeling.

Regarding the number of smoke constituents above the LOQ, the observed ranges reported in Table 5 were consistent with previously published levels of mainstream cigarette smoke (Counts et al., 2004, 2005).

3.2. Modeling smoke emissions

The computational process succeeded in building multiple linear regressions. The process was able to identify variables enabling the construction of simple modeling of smoke constituents under ISO 3308 conditions with six explanatory variables maximum while reaching high R-squared values (Fig. 1).

R-squared values obtained with models from one to six predictors to explain each of the 58 smoke constituents under ISO 3308 conditions are represented on Fig. 2. Dots correspond to a couple of metrics (number of predictor and R-squared). Boxplots provide statistical summaries computed from these metrics. The curve indicates the trend of R-squared versus the number of predictor variables. The R-squared values ranged from 0.442 to 0.913 for models involving one explanatory variable, whereas the range varied from 0.959 to 1.000 for models with six variables. Additionally, two variables were sufficient to model 84% (49 out 58) of smoke constituents under ISO 3308 conditions by reaching R-squared higher than 0.8. This indicates that multiple linear regression models succeeded in building simple models (i.e. with low number of variables) while explaining a large part of variations of smoke constituents under ISO 3308 conditions. It is important

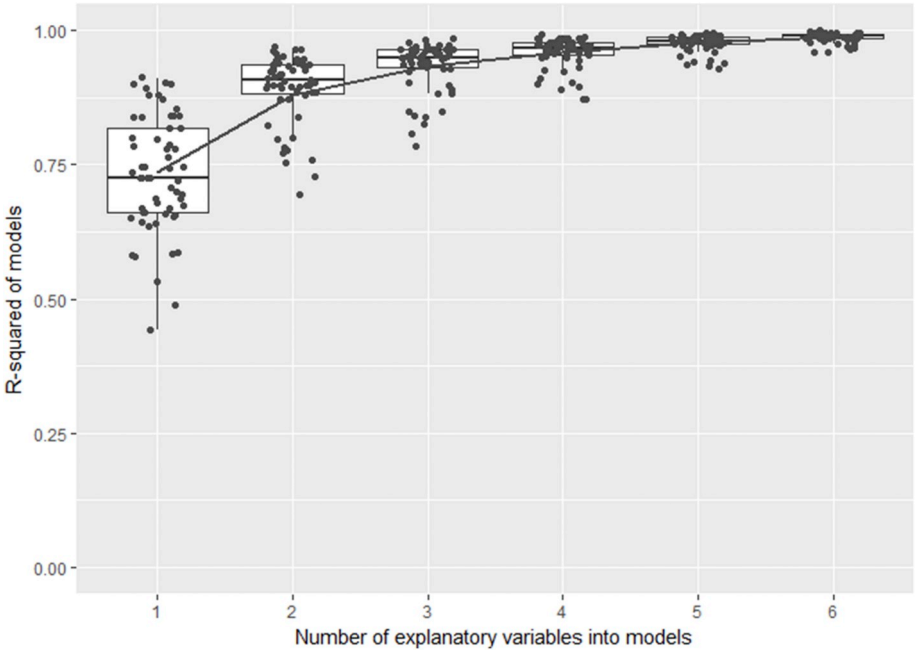

Fig. 2. Illustration of the number of explanatory variables used for modeling of smoke constituents under ISO 3308 conditions, and the corresponding R-squared.

Table 6
Contribution of the four blocks of information and potential explanatory variables never selected by the computational process for modeling smoke constituents under ISO 3308 smoking conditions.

Block n°	Block of information	Number of models using at least one variable *(on a total of 60 models)*	Number of available variables	Number of variables used in modeling	Potential explanatory variables never selected
1	Smoke constituents under intense conditions	60 *(100%)*	58	55 *(95%)*	NDMA, NNK, TPM
2	Tobacco filler constituents	19 *(33%)*	21	12 *(57%)*	Anabasine, Cadmium, Chromium, Chrysene, Cobalt, Lead, Mercury, Naphthalene, Nicotine
3	Cigarette design parameters	54 *(93%)*	14	11 *(79%)*	Butt length, Paper permeability, Smouldering rate
4	Cigarette burning parameters	25 *(43%)*	8	7 *(88%)*	Length of tobacco rod actively burnt

to emphasize that simple models are generally less sensitive and more robust because the risk of adding noise due to negligible variables is controlled, while also limiting the costs of data collection and model maintenance (Montgomery et al., 2001).

Table 5 describes the modeling of smoke constituents under ISO 3308 conditions by providing: the measured range; the R-squared corresponding to the ultimate best model selected according to the Bayes Information Criteria (BIC) and the number of variables used by the models. Out of the 58 smoke constituents, 57 (98%) have been modeled with coefficients of determination (R-squared) higher than 0.9. This includes 56 models (97%) with R-squared higher than 0.95 and 24 (41%) with R-squared higher than 0.99. The groups of phenolic and semi-volatile compounds had models with the highest range of R-squared (higher than 0.99). In addition, NPYR from the group of nitrosamines was modeled with the lowest R-squared (0.87) compared with the other smoke constituents under ISO 3308 conditions. This lowest R-squared might be explained by the lowest number of predictors selected by the BIC criteria, four compared to six for the others smoke constituents.

The high R-squared values obtained indicate that the variations of the smoke constituents under ISO 3308 conditions are properly explained by: smoke constituents emitted under intense conditions; tobacco filler constituents; cigarette design parameters and burning parameters. These findings are reinforced by the low number of explanatory variables used in the best regression equations. Indeed, four to six variables were sufficient to reach such high R-squared values.

Table 6 summarizes the contribution of the four blocks of information into the modeling of the 58 smoke constituents under ISO 3308 conditions. Smoke constituents under intense conditions (block n°1) are involved in the 58 modelings (100%) and constitute the main contributor. The second main contributor is the block n°3 (design parameters) involved in 54 modelings (93%). Then, blocks n° 4 (burning parameters) and n°2 (tobacco) contributed to 25 (43%) and 19 modelings (33%), respectively.

Although there is an important redundancy between smoke constituents under intense and less intense conditions, some additional predictor variables are required to enhance modeling. In particular, the filter ventilation (from block n°3) is involved in 39 (67%) modelings and constitute the most used variable. The weak contribution of the tobacco constituents (block n°2) can be explained by the fact that this block of information is partly included in the block n°1 (smoke constituents under intense conditions), through combustion and transfer, as published by Bry and Verron (2015). It shall be highlighted also that 16 (16%) variables out of the 101 available were never selected by the models (Table 6).

To summarize, the computational process used in this study succeeded to find subsets of variables to explain 98% of smoke constituents

under ISO 3308 conditions generated in the frame of this study. With the exception of NPYR and ammonia (R-squared values at 0.871 and 0.900 respectively) (Table 5), the additional information gained by adding a second smoking regime is no more than 5% (R-squared values higher than 0.95). This study supports the premise that the application of two smoking regimes is not useful and does not produce a significant product characterisation.

It should be noted that the number of products (22 in total) was limited. In order to confirm results, external validation and additional research should be conducted with a broader range of products. Furthermore, the correlation between smoke constituents for a given smoking regime was not analyzed in this study.

4. Conclusions

The objective of this study was to determine the added value of reporting smoke constituents under both intense and less intense conditions. A statistical process was developed to select subsets of features that are useful to build a good linear relationship. It is important to note that the objective of this study was to find all potentially useful variables but not necessarily the most relevant ones. Indeed selecting the most relevant variables is usually suboptimal for building a predictor, particularly if there is lot of correlations between variables. As such, the subsets of variables selected by our approach may exclude relevant variables. The good quality of adjustment (R-squared higher than 0.90 for 98% of smoke constituents) supports the usefulness of a second smoking regime. This study highlights, in the range of tested products, that an increase of the volume of data does not necessarily increase the volume of information and consequently does not necessarily improve knowledge. The debate should now be to determine which smoking regime to keep, considering the discrimination power of mainstream smoke constituents and a set of criteria such as: method variabilities, limits of detection and quantification and correlation between constituents.

Conflicts of interest

The authors would like to state that they have no competing interest. This research was funded by Imperial Tobacco Limited. The authors are employees of SEITA Imperial Brands.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.yrtph.2018.09.015.

Transparency document

Transparency document related to this article can be found online at https://doi.org/10.1016/j.yrtph.2018.09.015.

References

Akaike, H., 1998. Information theory and an extension of the maximum likelihood principle. In: Parzen, E. (Ed.), Selected Papers of Hirotugu Akaike. Springer New York, New York, NY, pp. 199–213.

Bry, X., Verron, T., 2015. THEME: THEmatic model exploration through multiple co-structure maximization. J. Chemometr. 29, 637–647.

Colard, S., 2015. Prediction model for cigarette yields derived from data obtained under two different machine smoking regimes. Beiträge Tabakforsch./Contrib. Tob. Res. 26, 320–333.

Colard, S., et al., 2014. Relationship between cigarette yields and smoking time under different machine smoking regimes. Beiträge Tabakforsch./Contrib. Tob. Res. 26, 4–18.

Counts, M.E., et al., 2004. Mainstream smoke constituent yields and predicting relationships from a worldwide market sample of cigarette brands: ISO smoking conditions. Regul. Toxicol. Pharmacol. 39, 111–134.

Counts, M.E., et al., 2005. Smoke composition and predicting relationships for international commercial cigarettes smoked with three machine-smoking conditions. Regul. Toxicol. Pharmacol. 41, 185–227.

Draper, N.R., Smith, H., 1981. Applied Regression Analysis. Wiley.

European Union Directive, 2001. 37/EC concerning the manufacture, presentation and sale of tobacco product. NOT, EN. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32001L0037.

Food and Drug Administration, 2012a. Draft Guidance; Reporting Harmful and Potentially Harmful Constituents in Tobacco Products and Tobacco Smoke under Section 904(a)(3) of the Federal Food, Drug, and Cosmetic Act. http://www.fda.gov/downloads/TobaccoProducts/GuidanceComplianceRegulatoryInformation/UCM297828.pdf.

Food and Drug Administration, 2012b. Harmful and Potentially Harmful Constituents in Tobacco Products and Tobacco Smoke: Established List. US Food and Drug Administration. http://www.fda.gov/TobaccoProducts/GuidanceComplianceRegulatoryInformation/ucm297786.htm.

Food and Drug Administration, 2016. Deeming tobacco products to Be subject to the federal Food, Drug, and cosmetic act, as amended by the family smoking prevention and tobacco Control act; restrictions on the sale and distribution of tobacco products and required warning statements for tobacco products. https://www.gpo.gov/fdsys/pkg/FR-2016-05-10/pdf/2016-10685.pdf.

Frost, J., 2015. The danger of overfitting regression models. The Minitab blog. http://blog.minitab.com/blog/adventures-in-statistics-2/the-danger-of-overfitting-regression-models.

Furnival, G.M., Wilson, R.W., 2000. Regressions by leaps and bounds. Technometrics 42, 69–79.

Hand, D.J., 1981. Branch and bound in statistical data analysis. J. R. Stat. Soc. - Ser. D Statistician 30, 1–13.

Hawkins, D.M., 2004. The problem of overfitting. J. Chem. Inf. Comput. Sci. 44, 1–12.

Health Canada, 1999. Method T-115: Determination of “Tar”, Nicotine and Carbon Monoxide in Mainstream Tobacco Smoke. Ottawa.

Health Canada, 2000. Tobacco Reporting Regulations SOR/2000-273.

ISO, 1999a. Standard 3402. Tobacco and Tobacco Products - Atmosphere for Conditioning and Testing. International Organization for Standardization, Geneva, Switzerland.

ISO, 1999b. Standard 10362-1. Cigarettes - Determination of Water in Smoke Condensates – Part 1: Gas-chromatographic Method, second ed. International Organization for Standardization, Geneva, Switzerland.

ISO, 2000a. Standard 3308. Routine Analytical Smoking Machine - Definition and Standard Conditions. International Organization for Standardization, Geneva, Switzerland.

ISO, 2000b. Standard 4387. Cigarettes – Determination of Total and Nicotine Free Dry Particulate Matter Using a Routine Analytical Smoking Machine. International Organization for Standardization, Geneva, Switzerland.

ISO, 2000c. Standard 10315. Cigarettes - Determination of Nicotine in Smoke Condensates – Gas-chromatographic Method, second ed. International Organization for Standardization, Geneva, Switzerland.

ISO, 2001. Standard 5725-6. Accuracy (Trueness and Precision) of Measurement Methods and Results. Part 6. Use in Practice of Accuracy Values. International Organization for Standardization, Geneva, Switzerland.

ISO, 2003. Standard 9512. Cigarettes - Determination of Ventilation - Definitions and Measurement Principles. International Organization for Standardization, Geneva, Switzerland.

ISO, 2009a. Standard 2965. Materials Used as Cigarette Papers, Filter Plug Wrap and Filter Joining Paper, Including Materials Having a Discrete or Oriented Permeable Zone and Materials with Bands of Differing Permeability - Determination of Air Permeability. International Organization for Standardization, Geneva, Switzerland.

ISO, 2009b. In: Standard 8454. Cigarettes - Determination of Carbon Monoxide in the Vapour Phase of Cigarette Smoke – NDIR Method, third ed. Amendment 1 International Organization for Standardization, Geneva, Switzerland.

ISO, 2013. Standard 2971. Cigarettes and filter rods - Determination of nominal diameter - Method using a non-contact optical measuring apparatus - Cigarettes et bâtonnets-filtres. International Organization for Standardization, Geneva, Switzerland.

King, B., et al., 2007. Mainstream smoke emissions of Australian and Canadian cigarettes. Nicotine Tob. Res. J. 9, 835–844.

Lumley, T., 2006. Leaps: Regression Subset Selection. R Package Version 2.9.

Montgomery, D.C., Peck, E.A., Vining, G.G., 2001. Introduction to Linear Regression Analysis, third ed. John Wiley & Sons, New York, NY.

Pazo, D.Y., et al., 2016. Mainstream smoke levels of volatile organic compounds in 50 U.S. Domestic cigarette brands smoked with the ISO and canadian intense protocols. Nicotine Tob. Res. J. 18 (9), 1886–1894.

Purkis, S.W., et al., 2013. Effect of puffing intensity on cigarette smoke yields. Regul. Toxicol. Pharmacol. 66, 72–82.

R Core Team, R., 2016. A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.

WHO Tobacco Laboratory Network, 2012. Standard operating procedure for method Intense smoking of cigarettes (SOP 01). http://www.who.int/iris/bitstream/10665/75261/1/9789241503891_eng.pdf?ua=1.

World Health Organization, 2008. The scientific basis of tobacco product regulation, second report of a WHO Study Group (TobReg). http://www.who.int/tobacco/publications/prod_regulation/trs_951/en/.

World Health Organization, 2015. The scientific basis of tobacco product regulation, fifth report of a WHO Study Group (TobReg). http://www.who.int/tobacco/publications/prod_regulation/trs989/en/.