Prediction of ignition delay using data-driven framework for straight chain alkanes

Auther-1,Auther-2,Auther-3

Indian Institute of Technology, Madras

Abstract

Ignition delay is important global combustion property. Ignition delay time(IDT) is generally measured using Shock-tube and RCM experiments. Calculation of IDT from simulation is computationally expensive and time consuming process. To obtain IDT faster and accurately, shock tube experimental data has been used to predict IDT using error based clustering regression. For that, IDT correlation is Arrhenius type in its nature which includes activation energy and reformulated using bond energy to avoid uncertainty and dependency of experimental parameters. To predict IDT, models are created using error based clustering algorithm. For that dataset is divided into recursively into sub-dataset using relative error between predicted and actual ignition delay using multiple regression and hypothesis testing. Result obtained using framework and correlation shows excellent agreement with experimental result.

Keywords:
Ignition delay prediction, Machine learning, Data-Driven ,Fuel, Error based clustering, IDT correlation, Framework

1. Introduction:

Combustion process is mainly characterized by transport processes and chemical reactions. When fluid undergoes chemical reaction, it liberates heat without external source of energy such sustainable process is called as Ignition. Ignition comprises series of coincidental physical and chemical processes which have different characteristic time scale, which is called as ignition delay.

Ignition delay gives important information about fuel reactivity. It is one of the major global physio-chemical combustion property. Ignition delay is mainly comprised of two parts: physical ignition delay and chemical ignition delay. Physical ignition delay depends on certain physical phenomena such as heating, fuel atomization, penetration of spray, and evaporation rate of fuel for different temperature range. Whereas, chemical ignition delay is mainly function of chemical characteristics of fuel, molecular structure, equivalence ratio, etc. Chemical ignition delay is main focus of this study which will be referred as ignition delay time(IDT or ID).

Ignition delay is crucial factor for design of combustor. Right amount of ignition delay is required for proper functioning of combustor. Ignition delay are generally calculated using reacting flow simulation which involves large number species and thousands of reaction including broad range of chemical and flow time scale. Calculation of IDT for various fuel and wide range of conditions is complicated and time consuming process. Ignition delay calculation using realistic-detail chemical mechanism requires full-scale numerical solvers which are computationally intensive.[1]

Taking motivation from such complications, acquired results gives simplified, accurate correlations along with efficient framework which is applicable over variety of fuels and wide range of conditions.

1.1. Literature review:

Substantial work has been done by researchers to calculate and correlate the ignition delay. Major ignition delay models are of Arrhenius-type.

Notations :

P = Pressure	T = Temperature
ϕ = Equivalence Ratio	τ = Ignition Delay Time
X_{Oxi} = Oxidizer mole fraction	X_{Fuel} = Fuel mole fraction
E_a = Activation Energy	δH = Change in enthalpy
R = Gas constant	A = Scaling factor
IDT = Ignition delay time	C = Constant

Horning et al. [2] has conducted study of different hydrocarbon at high-temperature to observe auto ignition and thermal decomposition. For n-alkanes at $\phi = 1$ follows obtained ignition delay correlation is given as-1,

$$\tau = 9.40 * 10^{-6} P^{-0.55} X_{Oxi}^{0.63} C^{-0.5} e^{46500/RT} \tag{1}$$

Where C is number of carbon atoms in the molecules. Correlation suggest that, ignition delay is not only function of pressure, temperature but it also depends on chemical characteristics of fuel. Such Arrhenius-type equations are constrained by certain physical condition. In parametric uncertainty [2], he has also mentioned that activation energy is very sensitive to ignition temperature.

In recent study of distillate fuels, **Fethi and Amir** [3] has obtained ignition delay correlation for gasoline and jet fuels using modified Arrhenius expression which applicable over wide range of conditions [P = 10-80 bar, ϕ = 0.5-2]. By statistical methods and numerical simulation obtained correlation are mentioned as [2],

$$\tau_{gasoline} = 6.76 * 10^{-}7 \frac{P}{20}^{-1.01} \phi^{1.13 - \frac{(17.59)}{T}} \exp \frac{29.39}{RT} \tag{2}$$

$$\text{for } T > \frac{1000}{-0.073 ln(\frac{P}{P_0}) + \phi^{-0.0338} + 0.0938} \tag{3}$$

$$\tau_{JetFuel} = 4.76 * 10^{-}7 \frac{P}{20}^{-1.21} \phi^{2.04 - \frac{(29.56)}{T}} \exp \frac{29.33}{RT} \tag{4}$$

$$\text{for } T > \frac{1000}{-0.0371 ln(\frac{P}{P_0}) + \phi^{-0.00727} + 0.0995} \tag{5}$$

Such correlations are useful for modeling of the fuel surrogates. But they are either fuel specific or constrained by physical conditions.

Goldsborough has also suggested traditional Arrhenius-based, power law formulation for iso-octane [4]:

$$\tau = A\phi^{\alpha}P^{\beta}X_{O_2}^{\gamma}exp(\lambda) \tag{6}$$

where, α, β, γ are third order polynomials, to capture changes in functionality across different ignition delay regimes. λ is overall activation energy which is expressed by addition of two quadratic expression which also includes pressure terms in it.

Zhao et al. [5] has developed two ignition delay model for hydrogen/air mixtures using High Dimensional Model Representation (HDMR). Piece-wise correlation gives alternative for full kinetic mechanism. For that 2000 data points were obtained by random sampling of uniform distribution from given range parameter.

$$\frac{1}{1600} < \frac{1}{T_0} < \frac{1}{800}, -1 < \log_{10} P < 2, \log_{10} 0.2 < \log_{10} \phi < 1 \tag{7}$$

Obtained conditioned were used to run SENKIN(chemkin) simulation to obtain ignition delay values. Thus complete data-set of 2000 data points was used to generate correlation. Rather than generating single correlation, 2000 data points were divided into six segments based on pressure and temperature. All piece-wise models gives $R^2 > 0.95$.

Upper NTC turnover points of n-butane, n-heptane and iso-octane exhibits typical kinetic and thermodynamic properties under different pressures. **Weiqi Ji et al.** [6] showed that, the ignition delay at the turnover states exhibits an Arrhenius dependence on the temperature and approximately inverse quadratic power law dependency on the pressure which implies that the temperature and pressure at turnover states are not independent and can be correlated as $\ln P \propto \frac{1}{T}$.

To predict auto-ignition and flame properties of multi-component fuel, **Neel et al.** [7] has used two machine learning algorithm random forest and neural network. For that, ignition delay data of toluene primary reference fuel [TPRF] extracted from kinetic simulation. While training the model, he has observed that the major concern of over-representative data as it does overfitting. Removal of under-representative data points predicted the IDT with better accuracy.

For prediction global combustion behavior, **Dussan et al.** [8] have used chemical functional group for analytical formulation of IDT. CH_2, CH_3 and benzyl-type functional group are used to represent n-alkyl, iso-alkyl and aromatics. High temperature formulation was generated using Scheffë simplex polynomial(first order) with natural logarithm, considering temperature dependency of each functional group's mass fraction. The low temperature correlation was obtained by third order polynomial. Study concludes that molecular structure governs the combustion chemical kinetic behavior.

Apart from formulation, clustering of data also plays important role in generation of appropriate ignition delay model. **Chinta et al.** [9] have proposed clustering of parameter rather than clustering of data. He has also used statistical testing for removal of redundant parameters. The obtained piece-wise models gives better control in non-isothermal continuous stirred tank reactor.

From above discussion, concluded major affecting parameters i.e temperature, pressure, fuel mass fraction and molecular structure is utilized for model development. Also taking motivation from piece-wise cluster, rather than generating single cluster, error-based recursive clustering tree has been implemented. **The goal** of present study is to find out correlation which generalize, efficient, applicable over wide range physical condition and variety of fuel. To obtain this, it is necessary to modify Arrhenius-type equation. Formulation of ignition delay is discussed in next

sections.

2. Ignition delay formulation:

Arrhenius-type ignition delay correlation has limitation of fuel specificity and are also bounded by constrain of affecting parameters. Major affecting parameters also carries uncertainty in results. To remove such complication, ignition delay correlation has to be reformulated.

In chemical reactions, the most sensitive parameter in ignition delay correlation is activation energy. Activation energy describes overall transformation of reaction and it only gives macroscopic information about reaction as intermediate species are not considered in any reaction. More microscopic information about the single step reaction can be obtained using the Eyring eqaution [10].

$$k = \kappa \left(\frac{k_B T}{h} \right) e^{\frac{\Delta S^{\ddagger}}{R}} e^{\frac{-\Delta H^{\ddagger}}{RT}} \tag{8}$$

Where, k_B = Boltzmann constant, T = absolute temperature, h = Planck's constant, $\Delta S^{\ddagger}$ = Activation Entropy, $\Delta H^{\ddagger}$ = Activation Enthalpy.

Eyring's equation is based on statistical and mechanical rationale of transition state theory whereas Arrhenius equation is empirical. These both equation are different in its nature. Relation between these two equations is possible when elementary reaction is as uni-molecular or bi-molecular. In such case, activation energy or energy barrier can be defined in terms of enthalpy of activation [10] which is closely related to bond energy.

$$E_a = \Delta H^{\ddagger} + nRT \tag{9}$$

According to transition state theory, when molecules with enough kinetic energy collides in certain orientation, it may generate activated complex. Bond structure of activated complex is different from reactants bond structure. $\Delta H^{\ddagger}$ plays critical role in bond formation or breakage. So, enthalpy of activated complex is related to enthalpy of reaction.

In combustion, heat of combustion ($\Delta H_{combustion}$) or heat of reaction ($\Delta H_{reaction}$ = -$\Delta H_{combustion}$) is directly related to bond dissociation energy [11] which can be expressed as,

$$\begin{aligned} \Delta H_{combustion} &= H_{reactants} - H_{products} \\ &\approx \text{Bond energy of Reactants} - \text{Bond energy of Products} \end{aligned} \tag{10}$$

Being point function, enthalpy of activation ($\Delta H^{\ddagger}$) for forward reaction can be expressed as difference between reactant state and activation state,

$$\Delta H^{\ddagger} = H_{reactant} - H^{\ddagger} \tag{11}$$

from equation-(10) & (11),

$$\Delta H^{\ddagger} = \Delta H_{combustion} + H_{products} - H^{\ddagger} \tag{12}$$

In IDT equation-(1) Horning et al. has showed that IDT depends on the number of carbons. So, Arrhenius-type ignition delay can be reformulated using equation-(9) and (12).

$$\begin{aligned}
\tau &= A \cdot \phi^{\alpha} \cdot P^{\beta} \cdot X_{O_2}^{\gamma} \cdot \exp\left(\frac{E_a}{RT}\right) \\
&= A \cdot \phi^{\alpha} \cdot P^{\beta} \cdot X_{O_2}^{\gamma} \cdot \exp\left(\frac{\Delta H^{\ddagger} + nRT}{RT}\right) \\
&= A \cdot \phi^{\alpha} \cdot P^{\beta} \cdot X_{O_2}^{\gamma} \cdot \exp\left(\frac{\Delta H_{combustion} + H_{products} - H^{\ddagger}}{RT}\right) \exp\left(\frac{nRT}{RT}\right) \\
&= A \cdot \exp(n) \cdot \phi^{\alpha} \cdot P^{\beta} \cdot X_{O_2}^{\gamma} \cdot \exp\left(\frac{\Delta H_{combustion} + H_{products} - H^{\ddagger}}{RT}\right) \\
&= C \cdot \phi^{\alpha} \cdot P^{\beta} \cdot X_{O_2}^{\gamma} \cdot \exp\left(\frac{\Delta H_{combustion} + H_{products} - H^{\ddagger}}{RT}\right)
\end{aligned} \tag{13}$$

From-(10), it clear that enthalpy of combustion depends on the type of bond and bond energy. Enthalpy being point function, it can be treated as constant term so, formulation can be rewritten as:

$$\begin{aligned}
\tau &\propto \phi^{\alpha} \cdot P^{\beta} \cdot X_{O_2}^{\gamma} \cdot \exp\left(\frac{\text{Fuel Bond Energy Term} + \text{Constant Energy Term}}{RT}\right) \\
&\propto \phi^{\alpha} \cdot P^{\beta} \cdot X_{O_2}^{\gamma} \cdot \exp\left(\frac{1}{T} \cdot \underbrace{\left\{\frac{\text{Constant Energy Term}}{R} + \frac{\text{Fuel Bond Energy Term}}{R}\right\}}_{term-A}\right)
\end{aligned} \tag{14}$$

Now, by dimensional analysis of Arrhenius exponent term,

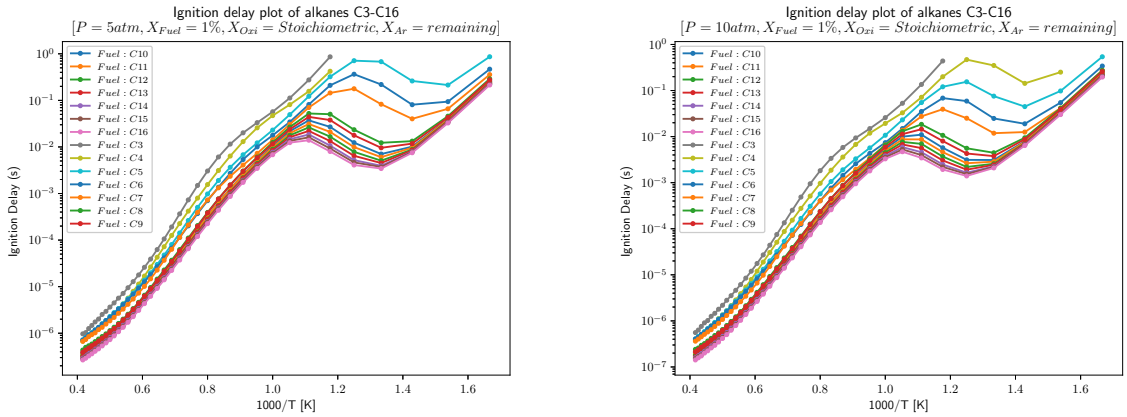
$$\begin{aligned}
\frac{E_a}{RT} &\sim \left[\frac{\frac{KJ}{mol}}{\frac{KJ}{mol \cdot K} \cdot K}\right] \\
\frac{E_a}{R} \cdot \frac{1}{T} &\sim \left[\frac{\frac{KJ}{mol}}{\frac{KJ}{mol \cdot K} \cdot K}\right] \\
&\sim \left[\frac{\frac{KJ}{mol}}{\frac{KJ}{mol \cdot K}}\right]\left[\frac{1}{K}\right] \\
&\sim \underbrace{\left[\frac{\frac{KJ}{mol} \cdot K}{\frac{KJ}{mol}}\right]}_{\text{Units of temprature}} \left[\frac{1}{K}\right] \\
&\sim \underbrace{(\alpha_0 \cdot T_0 + \alpha_1 \cdot T_0)}_{term-B} \cdot \frac{1}{T} \quad \text{where, } T_0 \text{ - Refrence Temprature}\cdot
\end{aligned} \tag{15}$$

Now, comparing term-A in (14) and term-B in (15),

$$\underbrace{\left\{\frac{\text{Constant Energy Term}}{R}+\frac{\text{Fuel Bond Energy Term}}{R}\right\}}_{term-A}\sim\underbrace{(\alpha_0\cdot T_0+\alpha_1\cdot T_0)}_{term-B} \tag{16}$$

From comparison, it is clear that β_1 contains information regarding types of bond. But it is difficult to obtain direct relationship between those parameters which can be obtained using chemical kinetic simulations of different fuels.

Chemical kinetic simulation of Propane-[12], Butane-[13], Pentane-[14], Hexane-[15], Heptane-[16], and Octane to Hexa-Decane-[17] fuel has been conducted using constant volume reactor at condition of P=5atm and 10 bar, $X_{Fuel} = 1\%$, X_{Oxygen}=According to stochiometry , X_{Ar} =remaining, T=600 to 2400 K by increment of 50K using ChemKin-18.1. From obtained result mentioned in fig-1(a) & 1(b) , it clear that from 2400 K (higher temperature) to around 1000K curve follows almost linear relationship. Slope of linear fit slope gives $\frac{E_a}{R}$.



(a) Ignition delay result of C3-C16 alkanes at 5atm (b) Ignition delay result of C3-C16 alkanes at 10atm

Figure 1: Ignition delay result of C3 - C16 alkanes at P= 5atm and 10atm, $X_{Fuel} = 1\%$, X_{Oxygen} =According to stochiometry, $X_{Ar} = Remaining$,T= 600 to 2400 K.

Obtained $\frac{E_a}{R}$ values for different fuels are shown in figure-3(a) and 3(b). All blue points shows $\frac{E_a}{R}$ value obtained by performing regression over simulation result. Plotting the $\frac{E_a}{R}$ against number of secondary carbon-hydrogen bonds-C_{SH} gives hidden information regarding relationship between two parameters. Procedure to find C_{SH} and other bonds information is mentioned in 3.1 and illustrated in fig- 5 by taking example of pentane.

In figure-2(a) and 2(b), points shows $\frac{E_a}{R}$ values obtained by regression considering data within temperature range of 1800 to 1250 K. Each data point of $\frac{E_a}{R}$ is associated with each fuel. Same way, points in figure-3(a) and 3(b) were obtained by performing regression considering all data from T= 600 to 2400 K. By looking at trend of data points, intuitively it looks obvious that $\frac{E_a}{R}$ data follows nearly inversely proportional relationship with C_{SH}.

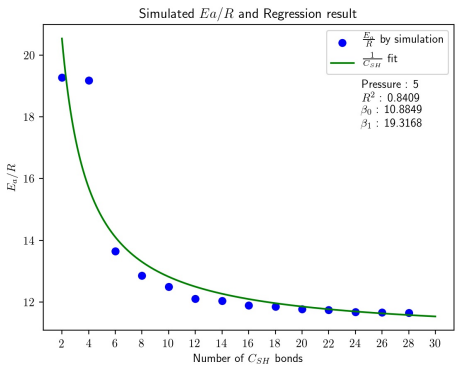
To verify the relationship between $\frac{E_a}{R}$ and C_{SH}- (Number of secondary carbon-hydrogen bonds) used hypothesis is mentioned below:

$$\begin{aligned} \frac{E_a}{R} &\propto \frac{1}{C_{SH}} \\ \frac{E_a}{R} &\propto \beta_0 + \beta_1 \cdot \frac{1}{C_{SH}} \\ \frac{E_a}{R} &\propto \beta_0 + \beta_1 \cdot {C_{SH}}^{-1} \end{aligned} \tag{17}$$

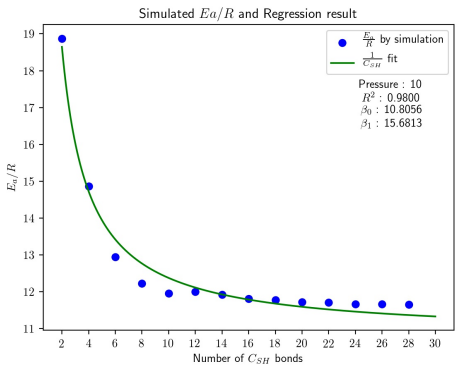
By inverting C_{SH} data and performing simple linear regression. Obtained regression coefficient and R^2 valued are mentioned in respective plots. All the values in fig-2(a), 3(a), 2(b), 3(b) shows promising results in favor of hypothesis with R^2 >0.84 for different cases.

To maintain dimensionality, temperature term is used as proportionality constant.

$$\frac{E_a}{R} = T_0 \cdot \beta_0 + \beta_1 \cdot \frac{T_0}{C_{SH}} \tag{18}$$

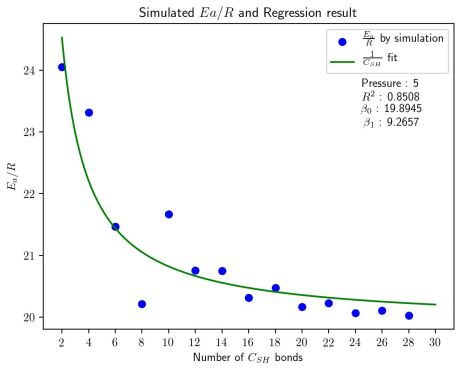


(a) E_a/R vs Number of C_{SH} at 5atm

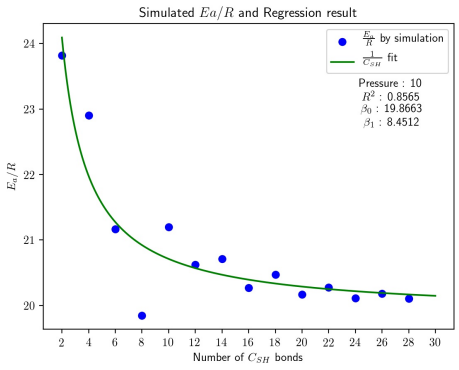


(b) E_a/R vs Number of C_{SH} at 10atm

Figure 2: E_a/R (using data ranges from 600 to 2400K) vs number C_{SH} bonds and also its regression fit



(a) E_a/R vs Number of C_{SH} at 5atm



(b) E_a/R vs Number of C_{SH} at 10atm

Figure 3: E_a/R (using data ranges from 1250 to 1800K) vs number C_{SH} bonds and also its regression fit

From equation- (16) and (18) the obtain term can be written as,

$$\left\{\underbrace{\frac{\text{Constant Energy Term}}{R}+\frac{\text{Fuel Bond Energy Term}}{R}}_{term-A}\right\}\sim\underbrace{(\alpha_0\cdot T_0+\alpha_1\cdot T_0)}_{term-B}\sim\underbrace{(T_0\cdot\beta_0+\beta_1\cdot\frac{T_0}{C_{SH}}}_{term-C}\Bigg\} \tag{19}$$

Replacing term-A of formulation-14 with term-C mentioned above, gives final formulation for ignition delay with replacement of activation energy with bond details - 20.

$$\tau\propto\phi^{\alpha}\cdot P^{\beta}\cdot X_{O_2}^{\gamma}\cdot\exp\left(\frac{1}{T}\cdot\left\{T_0\cdot\beta_0+\beta_1\cdot\frac{T_0}{C_{SH}}\right\}\right) \tag{20}$$

This formulation gives clear indication regarding, ignition delay dependency on the fuel bonds. To make quantities independent of unit, some terms in the formulation is normalized with unit quantities. Considering all the possible affecting parameters in above formulation-20, the functional form it can be expressed as:

$$\tau=f(T,P,\phi,X_{Fuel},X_{O_2},X_{Dilutant},C_{SH}) \tag{21}$$

But it is known that ,

$$X_{Fuel}+X_{O_2}+X_{Dilut6ant}=1\quad\&\quad\phi=\frac{\left(\frac{X_{fuel}}{X_{O_2}}\right)_{act}}{\left(\frac{X_{fuel}}{X_{O_2}}\right)_{stochio}} \tag{22}$$

As Fuel composition is already known, from stoichiometry equation, $\left(\frac{X_{fuel}}{X_{O_2}}\right)_{stochio}$ is attainable. It is possible to obtain ϕ from X_{fuel} and X_{O_2}, which is also true for $X_{Diluant}$.

Removing redundant function formulation can be re-written as,

$$\tau=f(T,P,X_{Fuel},X_{O_2},C_{SH}) \tag{23}$$

$$\begin{aligned}\tau&\propto\left(\frac{P}{P_0}\right)^{b}\cdot X_{Fuel}^{c}\cdot X_{O_2}^{d}\cdot exp\left(\beta_0\cdot\frac{T_0}{T}+\beta_1\cdot\frac{T_0}{T\cdot C_{SH}}\right)\\&=C'\cdot\left(\frac{P}{P_0}\right)^{b}\cdot X_{Fuel}^{c}\cdot X_{O_2}^{d}\cdot exp\left(\beta_0\cdot\frac{T_0}{T}+\beta_1\cdot\frac{T_0}{T\cdot C_{SH}}\right)\end{aligned} \tag{24}$$

Formulation can simplified by taking natural log on both side,

$$\begin{aligned}ln(\tau)&=ln(C')+b\cdot ln(\frac{P}{P_0})+c\cdot ln(X_{Fuel})+d\cdot ln(X_{O_2})+\beta_0\cdot\frac{T_0}{T}+\beta_1\cdot\frac{T_0}{T\cdot C_{SH}}\\ln(\tau)&=C+b\cdot ln(\frac{P}{P_0})+c\cdot ln(X_{Fuel})+d\cdot ln(X_{O_2})+\beta_0\cdot\frac{T_0}{T}+\beta_1\cdot\frac{T_0}{T\cdot C_{SH}}\end{aligned} \tag{25}$$

where, C, b, c, d, β_0, β_1 are coefficients, which is attainable from ignition delay data using multiple regression. It was assumed that ignition delay correlation depends on all the parameters which can be refined using hypothesis testing. To obtain models, multiple regression was used with error based clustering, which explained in further discussion. The fundamental step of whole process is to collect the data and make ot processable, which is discussed in next section.

3. Data collection and processing

Many researchers have performed experiments to measure ignition delay using different fuels for wide range of conditions. Focus of this study is to predict IDT of straight chain alkanes. The main source of data is Stanford group. Summary of ignition delay data obtained for wide range of conditions with variety of fuels is given in table-1:

Fuel		Temperature (K)	Temperature Uncertainty (%)	Pressure (atm)	Pressure Uncertainty (%)	Fuel Mole Fraction (%)	Oxygen Mole Fraction (%)	Equivalence Ratio	Data Points	Research Group	Reference
Propane	max	1841	±3	67.8	±1	4	20	5	174	Stanford	[18],[19],[20],[21],[22],[23]
	min	950	±0.7	1.12	±0.7	0.05	0.25	0.5			
Butane	max	1761	±3	5.5	±1	2	13	2	58	Stanford	[19],[18],[20]
	min	1230	±0.7	1.03	±0.7	0.05	0.325	0.5			
Pentane	max	1533	±0.7	3.75	±1	0.5	4	1	15	Stanford	[21],[24]
	min	1261	±0.7	1.62	±1	0.25	4	0.5			
Hexane	max	1475	±0.7	3.6	±1	0.42	4	1	16	Stanford	[21],[24]
	min	1237	±0.7	1.67	±1	0.21	4	0.5			
heptane	max	1784	±1.8	60.6	±1	1.874	20.6	2	107	Stanford	[18],[19],[20],[25],[26],[27],[28]
	min	806	±0.7	1.14	±0.7	0.03	0.33	0.5			
Octane	max	1455	±0.7	3.81	±1	0.32	4	1	15	Stanford	[21],[24]
	min	1252	±0.7	1.87	±1	0.16	4	0.5			
Nonane	max	1301	±0.7	41.76	±1	0.4	4	2	27	Stanford	[21],[24]
	min	1051	±0.7	13.52	±1	0.2	4	0.5			
Decane	max	1706	±2	5.15	±0.7	2.567	21	1.89	25	Stanford	[18],[19],[20],[29]
	min	1081	±1.5	1.22	±9.6	0.03	0.3875	0.64			
Dodecane	max	1657	±1	33.7	±9.3	2.138	21	1.88	162	Stanford	[30],[31],[32],[33],[34],[35],[36]
	min	727	±0.7	2.07	±0.7	0.0371	0.731	0.05			
Hexadecane	max	1355	±2	6.77	±0.7	0.1832	4	1.22	18	Stanford	[37],[38],[29]
	min	1159	±2	1.71	±0.7	0.0312	1	0.56			

Table 1: Summary of ignition delay data with physical parameter and uncertainty in it

Stanford group has performed IDT experiment for short to long chain alkanes using different shock-tubes. In experiments, IDT is mainly obtained using species profile of OH, CO_2, CH, and CH_3 which is also a more accurate way to measure IDT. While measuring IDT from shock-tubes, inherent and inevitable phenomena generates uncertainty by boundary layer growth and causes reduction in amplitude of shocks. Approximate values of uncertainty is also reported in the Table-1 which can be also useful to generate synthetic data points within error bound. Gauthier et al.[25] reported that ignition delay time obtained by experiments have average uncertainty of $\pm 15\%$, which may actually vary from $\pm$ 10% to $\pm$ 30 %.

For certain data points, temperature and pressure uncertainty values were not reported. Those were replaced by the least available uncertainty value from whole dataset. It is clear from the dataset that, extensive study of Propane, Heptane, Dodecane has made availability lot of data points whereas Pentane, Hexane, Nonane and Decane has less experimental data points. Such imbalance dataset may cause bias in the result while performing multiple-linear regression. Such issues were resolved by generating more data points from uncertainty value of each parameter. Procedure is explained in further discussion.

As mentioned earlier, apart from physical parameters, information related to bond play key role for IDT correlation. In next section, procedure to obtain fuel bond information is discussed.

3.1. Fuel Bond Information

Type of bonds plays key role in combustion process as chemical transformation and energy release is only possible through bond breakage. Different fuels have varying bond type. Number of such bonds also varies within same group of fuel.

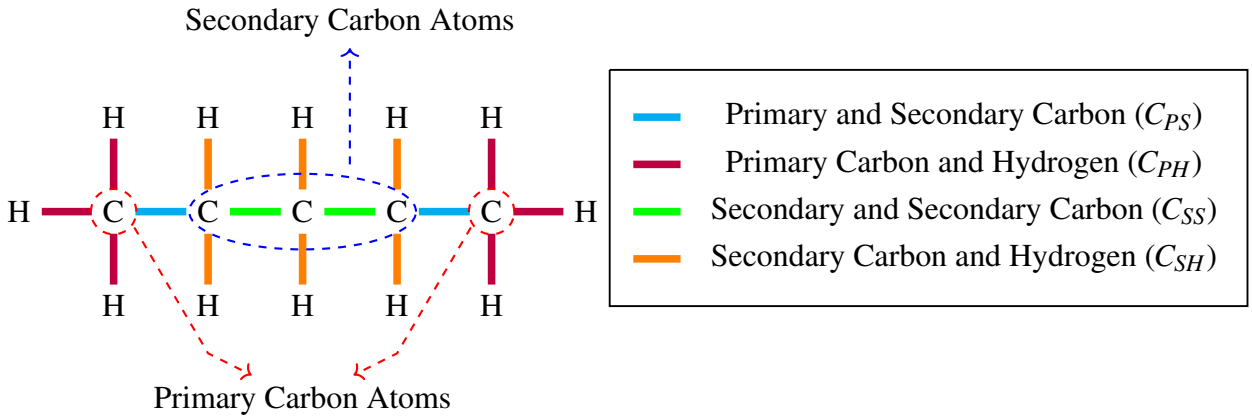


Figure 4: Different type of bonds in straight alkane (Pentane)

Bond information of fuel is obtained using SMILES (Simplified molecular-input line-entry system), which is a method for computer to convert chemical structure into processable input to extract the information. SMILES is based on molecular graph theory. Certain rules[39] are defined to transform the chemicals into the SMILES.

RDKit [40] famous python based library to process all SMILES and extract information associated with chemical structure and bond details. Using SMILES and RDKit, Mol-file was obtained. Mol-file gives information related to bond connections and orientation of bond in 3D space. By text processing of Mol-Block(file) bond information was extracted.

Type of bond in alkane is illustrated by taking example of pentane-5. It is clear from the figure that, straight chain alkanes (for carbon chain > 2) contains 4 different type of bonds. Ethane is only fuel which has bond type of primary-primary carbon so that fuel is excluded from the analysis. Extracted bond by processing smile and other parameters complete dataset is generated which is given in table-2. It is also observed that C_{PS}, C_{SS},C_{PH} and C_{SH} linearly dependent to each other. After generating complete dataset, processing of data is explained explain in further discussion.

Fuel	T(K)	T_Error(%)	P(atm)	P_Error(%)	Fuel(%)	Oxygen(%)	C_{PS}	C_{SS}	C_{PH}	C_{SH}	Time(μs)
CCC	1376	±1	1.19	±1	4	20	2	0	6	2	357
CCCC	1409	±1	1.17	±1	1	6.5	2	1	6	4	390
CCCCC	1395	±0.7	3.47	±1	0.5	4	2	2	6	6	316
CCCCCC	1273	±0.7	3.32	±1	0.42	4	2	3	6	8	1046
CCCCCCC	1378	±1	2.326	±1	0.03	0.33	2	4	6	10	1330
CCCCCCCC	1289	±0.7	2.01	±1	0.32	4	2	5	6	12	1198
CCCCCCCCC	1107	±0.7	14.8	±1	0.4	4	2	6	6	14	965
CCCCCCCCCC	1081	±2	5.14	±9.6	1.44	21	2	7	6	16	1696
CCCCCCCCCCC	1045	±1	6.71	±9.3	0.6098	21	2	9	6	20	2753
CCCCCCCCCCCCCC	1181	±2	2.13	±1	0.1776	4	2	13	6	28	5536

Table 2: Complete fuel dataset obtained after processing bond information. Fuel is represented by SMILES. T_Error(%) and P_Error(%) represents uncertainty in measurement of temperature and pressure respectively. Fuel bond notation are same as mentioned in figure-5

.

4. Data Clustering and Analysis:

The complete data-set of alkane fuel contains ignition delay as dependent variable and temperature, pressure, fuel and oxygen mole fractions as predictors. In inceptive period of model development, multiple linear has been used without any clustering technique. Obtained result had poor coefficient of determination around $R^2 \sim 0.70$ which indicates that model failed to capture proper variability.

Ignition delay data of alkanes have 5 independent parameters and different type of fuels. Model development of such data-set is not straightforward as it contains variety of fuels. As mentioned earlier, **Weiqi Ji et al.** [6] has divided the data based on the turnover states to revels distinct kinetic and To find IDT correlation of hydrogen/air mixture, **Zhao et al.** [5] has divided the data into six sub domains based on high, intermediate and low pressure along with high and low temperature range. Such approachhigh of grouping the data based on parameter indicates that ignition delay data has certain clustering pattern associated with it. For model development, along with regression, clustering also plays essential role.

It is possible to unveil the concealed pattern of n-dimensional data using clustering technique. In this study, new algorithm has been proposed to generate the clusters based on relative error of prediction.

4.1. Clutering based on relative prediction error:

The proposed clustering algorithm depends on the relative error between predicted and actual value of the dependent variable.

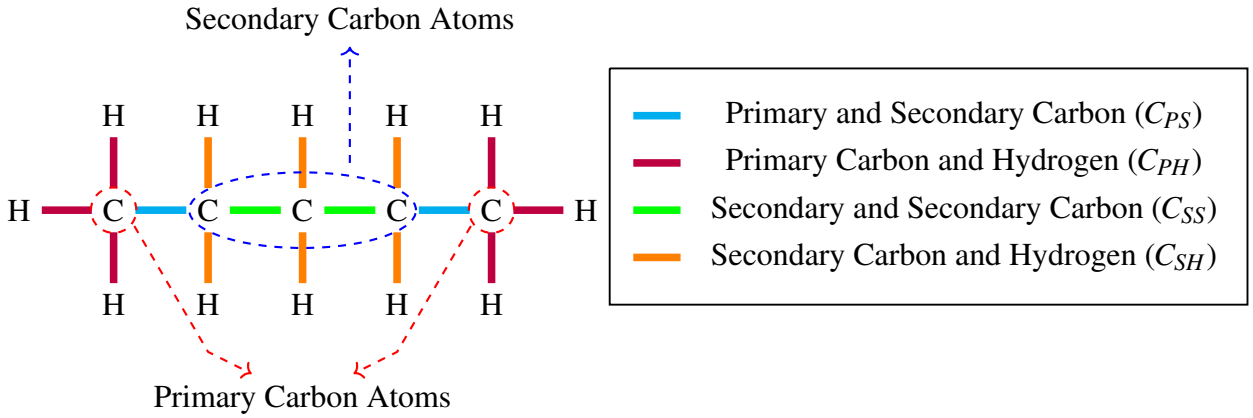


Figure 5: Different type of bonds in straight alkane (Pentane)

Number of clusters will be decided based on the accuracy of prediction and available data. It performs better in terms accuracy and execution time compared to other algorithm[41]. K-Means algorithm generates K centroids and minimizes the distance between cluster centroids and data-points for given number of clusters.

K-Means Algorithm:

- Consider the data points $x_i = x_1, x_2, x_3, ..., x_n \in \mathbf{R}^d$ has j features. $\mu_i \in \mathbf{R}^d$ denotes number of clusters and $q_1, ..., q_n \in \{1, ..., K\}$ denotes points assigned to the centroid. Such that the sum

of distances is minimized within cluster which is defined as,

$$E(\mu_1, \mu_2, .., \mu_K, q_1, q_2, ..., q_n) = \sum_{i=1}^{k} ||x_i - \mu_{q_i}||_P^p \tag{26}$$

where p denotes the norm.

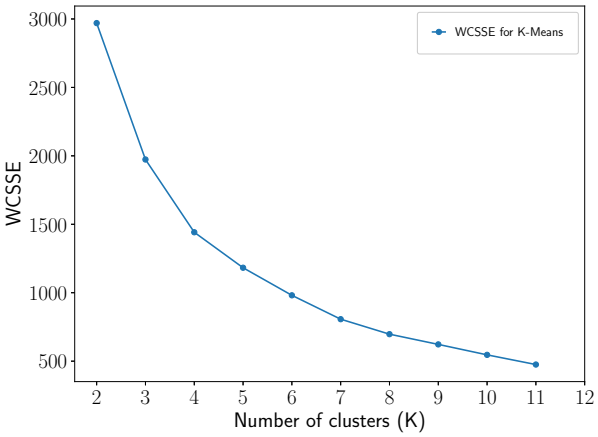
- steps:
  1. Randomly select the K initial cluster centroid.
  2. Assign the data points to nearest cluster
  3. Recompute the cluster based current data members of associated cluster
  4. Repeat the procedure till convergence criteria not met or clusters are not moving (Go to step:2)

K-Means algorithm is used from scikit-learn library. Which computes the clusters using Elkan's algorithm [42]. K-Means is useful method when number of clusters are known. But for the case of IDT, number of clusters are unkowns. To find out optimal number of clusters, elbow method is used along with silhouette criteria.

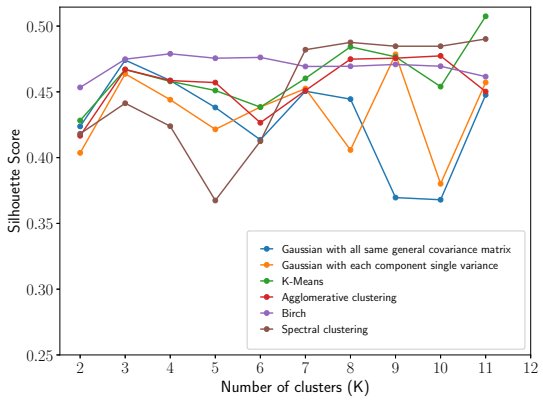
Elbow method calculates within-cluster sum of squared error (WCSSE) for given k cluster by equation-27. The K-value for which WCSSE starts declining, gives the optimized number of clusters as it minimizes the WCSSE.

$$WCSSE = \sum_{i=1}^{n} \sum_{x \in S_i} ||x_i - \mu_i||^2 \tag{27}$$

For ignition delay data, WCSSE is calculated for 2 to 11 number of clusters. Apart from ethane, all available data has been used to generate the clusters. From fig-6(a), it was observed that elbow(sharp slope) was generated around K=3 or K=4. The obtained result from fig-27 is quite ambiguous to analyse. To have more robust decision Silhouette criteria was needed along with elbow method.



(a) Elbow Method Result - Within-cluster sum of squared error(WCSSE) vs Number of cluster to decide the optimum number of cluster for K-Means



(b) Silhouette Score vs Number of cluster for different clustering technique

Figure 6: Criteria to decide the number of clusters for all fuel components

Silhouette Score Calculation:[43] [44]

- Data were divided into K-clusters
- Each data point x_i is assigned to cluster C_k such that ,$\forall x\ x_i \in C_k$
- Silhouette criteria measures the relative separation distance between points in same cluster to other clusters. Mean distance **within cluster** from point i to other points were measured by,

$$a(i) = \frac{1}{|C_i|-1} \sum_{j \in C_i, i \neq j} d(i,j) \tag{28}$$

  where, $d(i,j)$ is the distance between data points i and j in the cluster C_i. Numerator contains $|C_i|-1$ as $d(i,i)$ is excluded. In summary, $a(i)$ shows how well data point i is assigned to its cluster. Smaller value is expected for proper assignment.

- Mean distance between point i in cluster C to **points in other cluster** were measured by following equation:

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j) \tag{29}$$

  Use of min operator on cluster k shows that out of all value of different cluster, minimum value is used for calculation of $b(i)$ which is nearest neighbouring cluster to that specific point-i. Large value of $b(i)$ suggest that point-i does not properly matches with neighbouring cluster.

- Silhouette for points-i is defined as,

$$s(i) = \begin{cases} \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} & |C_i| > 1 \\ 0 & |C_i| = 1 \end{cases} \tag{30}$$

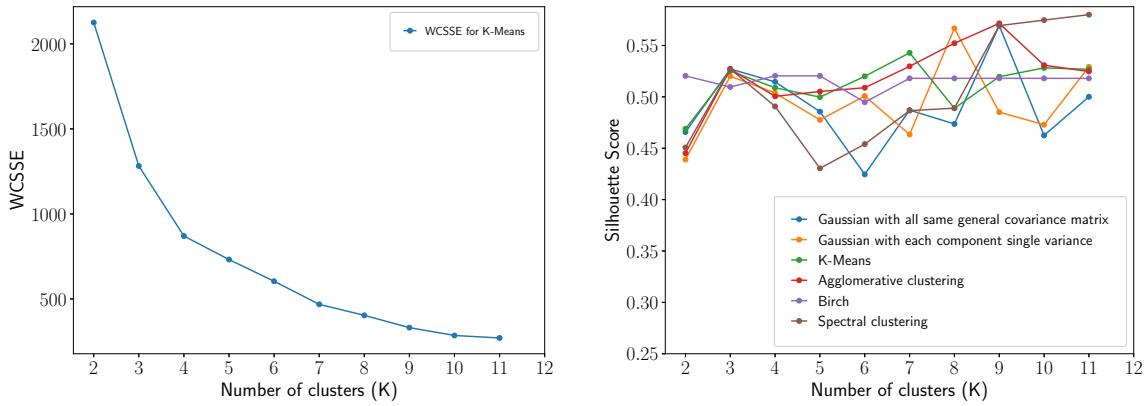
which is also defined as,

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & if \quad |a_i| < |b_i| \\ 0 & if \quad |a_i| = |b_i| \\ \frac{b(i)}{a(i)} - 1 & if \quad |a_i| > |b_i| \end{cases} \tag{31}$$

- $s(i)$ was calculated for all the points. The average value $s(i)$ in the cluster shows how tightly all point are assembled in the cluster. Mean $s(i)$ of of whole dataset shows how properly data points forms the group.

Silhouette analysis is useful for each data point to measure the distance between the clusters they belong and other neighbouring clusters. Silhouette Score varies from -1 to 1 in which 1 is ideal which shows that object are well separated and -1 shows improper clustering. To calculate Silhouette Score, sklearn.metrics.silhouette_score function is utilized.

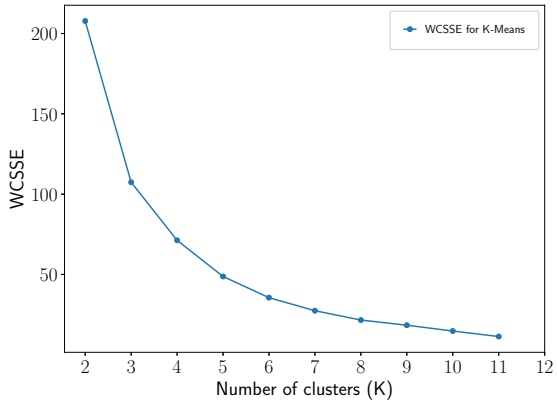
For IDT dataset, Silhouette analysis has been done using different clustering algorithm. First drop in result was observed at cluster K=3 for clustering techniques which emphasis that data should be divided into three partition. IDT dataset-1 has large variation in data points. Observed result is quite sensitive to type of fuel, conditions affecting to IDT and number of data points. For verification of technique and to check biasness in the result, silhouette analysis has been done on fuel which has more than 100 data points. From table-1, it observed that propane, heptane and dodecane satisfies that criteria. Those fuel also covers wide of range of conditions. The observed result given in fig-7(a) and fig-7(b) shows similar trend obtained in case if 'all fuel' - 6(b) 6(a). It clear from the fig-7(b) that number of cluster required for dataset of propane, heptane and dodecane is also 3.



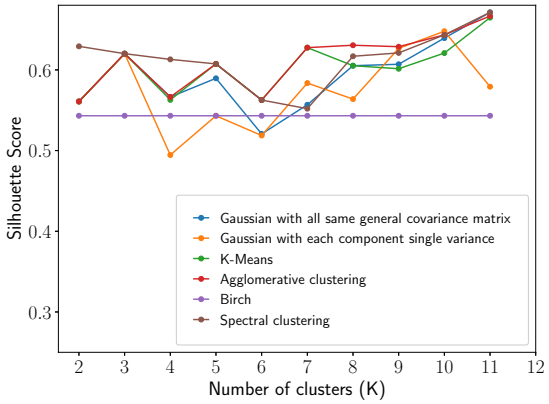
(a) Elbow Method Result: Within-cluster sum of squared error(WCSSE) vs Number of cluster for K-Means using data points of Propane, Heptane and Dodecane Fuel

(b) Result of Silhouette Score vs Number of cluster for different clustering technique using data points of Propane, Heptane and Dodecane

Figure 7: Criteria to decide the number of clusters obtained using Propane, Heptane and Dodecane



(a) Elbow Method Result: Within-cluster sum of squared error(WCSSE) vs Number of cluster for K-Means using data points of Propane

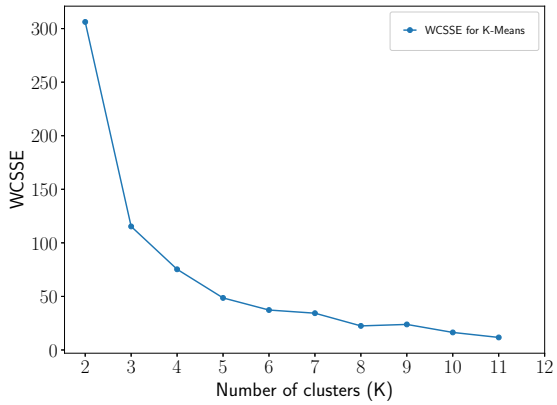


(b) Result of Silhouette Score vs Number of cluster for different clustering technique using data points of Propane

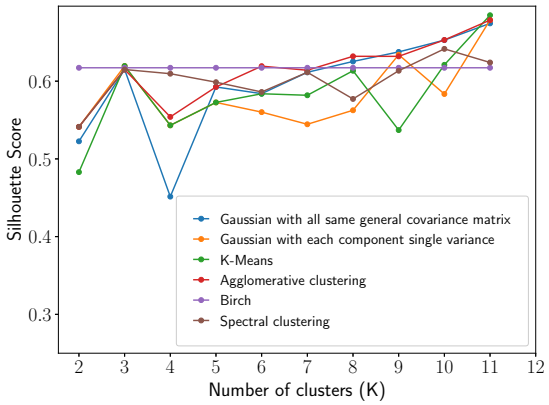
Figure 8: Criteria to decide the number of clusters obtained using Propane

Similar result was also observed in case of individual fuel dataset of Propane, Heptane as given in figure-8(b),9(b) which also agrees with expected trend. For propane and heptane observation matches with expected trend whereas in case of heptane -25(b), it suggest to divide the dataset into two sub-dataset as Dodecane data points has limited range of physical conditions. Dodecane data, in table-1 shows that, compared to heptane and propane data is limited by pressure at 33.7 whereas, propane and heptane has pressure around 60 atm. Similar observation is also made in equivalence ratio.

Silhouette score obtained in all the cases is above 0.4 which emphasis that data points are grouped by certain parameters. At a same time, clustering value is far away from 1 which means that they do not generate perfect cluster and data are quite separated which is useful information for regression analysis.

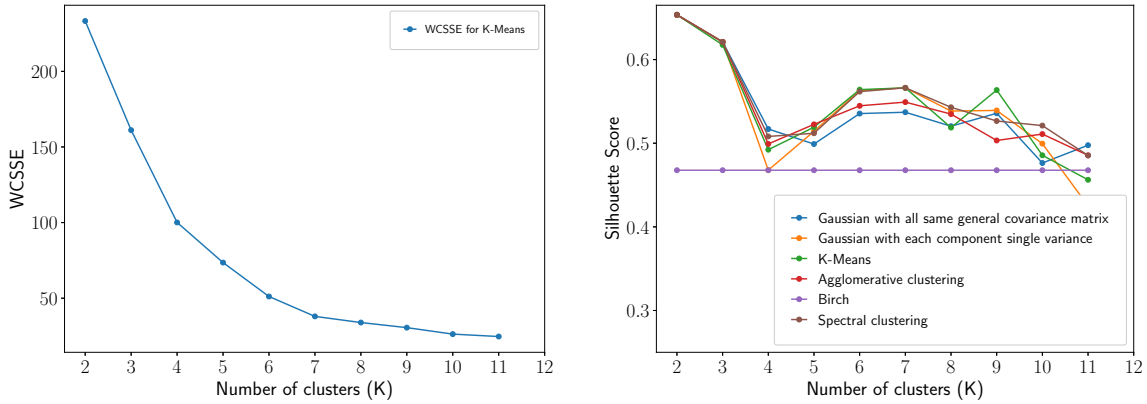


(a) Elbow Method Result: Within-cluster sum of squared error(WCSSE) vs Number of cluster for K-Means using data points of Heptane



(b) Result of Silhouette Score vs Number of cluster for different clustering technique using data points of Heptane

Figure 9: Criteria to decide the number of clusters obtained using Heptane



(a) Elbow Method Result: Within-cluster sum of squared error(WCSSE) vs Number of cluster for K-Means using data points of Dodecane

(b) Result of Silhouette Score vs Number of cluster for different clustering technique using data points of Dodecane

Figure 10: Criteria to decide the number of clusters obtained using Dodecane

4.2. Principal Component Analysis (PCA) of Dataset

Principal component analysis is useful for dimensionality reduction and for analysis of major component of the data. For given data, PCA find out major eigen values and eigen vector. Principle components obtained using full SVD solver showed that, data has 6 major eigen values and other two are near to 0. For (616, 8) size of dataset obtained variation in principal components(Eigen values) are obtained as below:

$$\begin{array}{cccc} 4.80958553e+00 & 2.55919622e+00 & 1.15728172e+00 & 9.98961156e-02 \\ 2.33198384e-02 & 1.68372670e-04 & 2.93116344e-32 & 1.18120890e-33 \end{array}$$

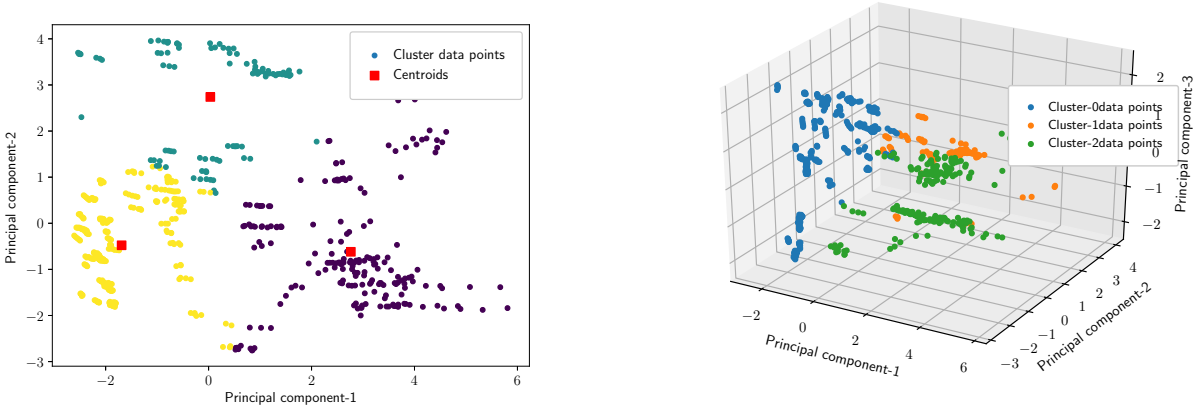
From six eigenvalues it is clear that, it has major three principal eigen values and out of rest, other three are smaller eigen values. so, it is possible to visualize the major variation in data and clustering in the 3-Dimensions. To visualize the clustering of data in 3D, K-Means along with PCA was implemented. Obtained result are given in fig-11(a) and -11(b). Obtained result also supports the the criteria to divide the dataset into 3 component.

4.3. Analysis of classified sub-dataset

Main dataset is divide into the three sub-dataset using silhouette criteria and K-Means clustering algorithm. By analysing the sub-dataset, obtained observations are discussed here.

4.3.1. Subdataset-I

It contains 331 data points. Mostly short and middle level alkanes at high pressure and temperature are part of this dataset. Certain data points at low temperate and high pressure are also part this set i.e. propane at 950K and 21.4 atm.



(a) Major 2 components ontained by PCA and K-Menas on that major 2 components

(b) Major 3 components ontained by PCA and K-Menas on that major 2 components

Figure 11: PCA and K-Means of all the fuels

Fuel		Temperature (K)	Pressure (atm)	Fuel Mole Fraction (%)	Oxygen Mole Fraction (%)	Equivalence Ratio	Data Points
Propane	max	1841	67.8	4	20	5	169
	min	950	1.12	0.15	0.75	0.5	
Butane	max	1761	5.5	2	13	2	55
	min	1230	1.03	0.5	3.25	0.5	
Pentane	max	1533	3.75	0.5	4	1	15
	min	1261	1.62	0.25	4	0.5	
Hexane	max	1475	3.6	0.42	4	1	16
	min	1237	1.67	0.21	4	0.5	
Heptane	max	1676	16.72	1.874	20.6	2	66
	min	1048	1.14	0.36	2.2	0.5	
Octane	max	1455	3.81	0.32	4	1	10
	min	1265	1.87	0.32	4	1	

Table 3: Summary of sub-data:II with label 0 using K-Means algorithm

4.3.2. Subdataset-II

It contains 92 data points. it almost covers all type of alkanes(by length). Mainly data points are separated by low fuel and oxygen mole fraction percentage.

Fuel		Temperature (K)	Pressure (atm)	Fuel Mole Fraction (%)	Oxygen Mole Fraction (%)	Equivalence Ratio	Data Points
Propane	max	1687	2.24	0.05	0.25	1	5
	min	1505	2.13	0.05	0.25	1	
Butane	max	1761	2.16	0.05	0.325	1	3
	min	1531	2.02	0.05	0.325	1	
Heptane	max	1784	15.81	0.2	2.2	1	24
	min	1229	1.61	0.03	0.33	0.5	
Octane	max	1434	2.19	0.16	4	0.5	5
	min	1252	1.91	0.16	4	0.5	
Decane	max	1706	2.28	0.2	3.1	1.2	20
	min	1327	1.22	0.03	0.3875	0.64	
Dodecane	max	1657	15.72	0.05146	1.911	1	31
	min	1252	2.07	0.0371	0.731	0.5	
Hexadecane	max	1333	6.77	0.0497	1	1.22	4
	min	1170	4.44	0.0371	1	0.76	

Table 4: Summary of sub-data:I with label 1 using K-Means algorithm

4.3.3. Subdataset-III

It contains 92 data points. Form intermediate to long chain alkanes at slightly lower temperature and high pressure along with high fuel and oxygen mole fractions are part of these dataset. Mainly it contains long chain alkanes with high fuel and oxygen content compared to other dataset.

Fuel		Temperature (K)	Pressure (atm)	Fuel Mole Fraction (%)	Oxygen Mole Fraction (%)	Equivalence Ratio	Data Points
Heptane	max	1115	60.6	1.874	20.6	1	16
	min	806	18.13	1.874	20.6	1	
Nonane	max	1301	41.76	0.4	4	2	27
	min	1051	13.52	0.2	4	0.5	
Decane	max	1173	5.15	2.567	21	1.89	5
	min	1081	4.56	1.44	21	1.06	
Dodecane	max	1422	33.7	2.138	21.0	1.88	131
	min	727	4	0.0558	2.786	0.05	
Hexadecane	max	1355	6.46	0.1832	4	1.22	14
	min	1159	1.71	0.0909	4	0.56	

Table 5: Summary of sub-data:III with label 2 using K-Means algorithm

5. Synthetic data generation and sampling technique:

To attain the objective, sufficient data points are necessary. From table- 3, 4, 5, it is clear that dataset is imbalance in terms of number of points for certain fuel. Coefficients obtained using multi-linear regression of such dataset will directly bias towards the more available information.

To reduce the bias in the result, different sampling strategy are utilized. This problem can be solved by two ways: [45]

1. Under-Sampling : Down sizing the largest dataset to minority size.
2. Over-Sampling : Expanding minority dataset size to majority dataset size

In under-sampling, negligence of data affect the correlation as less points may only cover limited range of physical condition. Due to drop of information, correlation may become detail specific. In over-sampling, expansion of data points is done by replicating the same data points which may ignore important information while sampling and may cause over-fitting. For IDT correlation rather than replicating samples uncertainty information is used to for over-sampling.

From table-1 it is clear that every fuel data-point contains certain range of error in measurement of temperature and pressure. Experimental error are mainly of two kinds:

1. Random Error : Occurs due to unknown and predictable changes
2. Systematic Error : Occurs due to measuring instrument and system handling

Error in measurement of shock temperature and pressure is uncontrolled phenomena which causes random error. Random error generally follows the Gaussian distribution. Using this information, more data points are generated using Multivariate Gaussian Normal Distribution.

$$p(x:\mu,\sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} \cdot |\sum|^{\frac{1}{2}}} \cdot exp\left(\frac{1}{2}(x-\mu)^T \sum^{-1} (x-\mu)\right) \tag{32}$$

where,
$\sum$ is covariance matrix between parameter.
μ is vector of mean values or reported value of parameter
n is number of parameters

5.1. Algorithm to generate synthetic data points

Algorithm 1 Algorithm to generate random samples using uncertainty

NUF = Number of Unique Fuel in the dataset
MNDP = Maximum Number of Data Points in largest fuel set
DPSF = total number of Data Points in Selected Fuel
TNDP = Total Number of Data Points in Extended dataset
DPG = number of Data Points to be Generated from each sample

TNDP= NUF * MNDP
for `i =1,2,3,..,NUF ∈ fuel data points` **do** ▷ Select Unique Fuel
    DPG = TNDP / DPSF ▷ P2oints from each sample
    for `j=1,2,...,DPSF` **do**
        `DATA_POINT_GENERATOR(data_point,DPG)` ▷ Append to dataframe
return Extended Dataframe ▷ Contains all fuel data

procedure DATA_POINT_GENERATOR(data_point,DPG)
    μ = [T_means, P_mean, τ_{mean}] ▷ Reported values in the dataset
    $\sum = \begin{bmatrix} \sigma_T^2 & 0 & 0 \\ 0 & \sigma_P^2 & 0 \\ 0 & 0 & \sigma_\tau^2 \end{bmatrix}$ ▷ From error values
    Generate 2000 points Using Multivariate Normal Gaussian $\mathcal{N} \sim \mathcal{N}(\mu, \sum)$
    Random sampling by DPG count ▷ to equi-size the dataset
    return Extended Dataframe ▷ extended from single data point

To generate more data points using error values, reported value of temperature and pressure is used as mean values and given error range considered as standard error. Standard error values taken as standard deviation to maximize the range of uncertainty. Fro single point n is considered as 1.

$$\sigma_{SD} = \frac{\sigma_{error}}{\sqrt{n}} \approx \sigma_{error} \tag{33}$$

Out of 9 dimensions, only 3 dimensions are considered to generate data points as other parameters are constant. Constant parameters were copied as it is in the data-frame as only temperature, pressure and IDT carries uncertainty. Reported value pressure and temperature is used for standard deviation. For IDT uncertainty varies maximum $\pm 30\%$ but to be within bound, $\pm 20\%$ uncertainty is used. Using multivariate Gaussian distribution along with relevant $\sum$ and μ, 2000 random points are generated to cover maximum possible occurrence. From 2000 random data points, sampling is done in such way that it will normalize the size of data points for all fuel. Procedure is mentioned in algorithm-1. To generate 2000 data points $np.random.multivariate\_normal(mean, cov, 2000).T$ function used and for sampling $random.choices(data, k = data\_generation\_count)$ is utilized.

Once extended equi-sized data-frame is generated it data should be divided into 3 sub-dataset using K-means algorithm and then individual data-frame is transferred to Multi-linear regression module.

6. Multiple Regression : OLS estimator and hypothesis testing

Multi linear regression model is given by.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i \tag{34}$$

where,
$Y_i = i^{th}$ observation of dependent variable
$X_{1i}, X_{2i}, X_{3i}, ..., X_{ki}$ are independent observation of the k regression
ε_i is the error term which is not covered by independent variables
Using ordinary least square, estimation of coefficient is done using observed values.

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + ... + b_k X_{ki} \tag{35}$$

where,
$b_0, b_1, b_2, ..., b_k$ are the estimated value of the $\beta_0, \beta_1, \beta_2, ..., \beta_k$ and coefficient $b_0, b_1, b_2, ..., b_k$ are obtained by solving linear system of equation.

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & X_{13} & ... & X_{1k} \\ 1 & X_{21} & X_{22} & X_{23} & ... & X_{2k} \\ 1 & X_{31} & X_{32} & X_{33} & ... & X_{3k} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & X_{n3} & ... & X_{nk} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}$$

For IDT dataset, dependent and independent variables are mentioned in formulation-23. In IDT equation-25, it is assumed that ignition delay depends on all the parameters and obtained from over-determinant system (n > k) of equation using least-square solution. Objective function is defined as,

$$\hat{b} = \underset{b}{\text{argmin}} \quad ||y - Xb||_2 \tag{36}$$

where, $\hat{\beta}$ is coefficient for best fit rather than the actual solution.

$$\begin{aligned} \hat{b} &= \underset{b}{\text{argmin}} \quad ||y - Xb||_2 \\ &= \underset{b}{\text{argmin}} \sum_{i=1}^{n} |y_i - \sum_{j=1}^{k} X_{ij} b_j| \\ &= \underset{b}{\text{argmin}} \quad (y - Xb)^T (y - Xb) \\ &= \underset{b}{\text{argmin}} \quad (y^T - b^T X^T)(y - Xb) \\ &= \underset{b}{\text{argmin}} \quad y^T y - y^T Xb - b^T X^T y + b^T X^T Xb \end{aligned} \tag{37}$$

Differentiating with respect to b^T ,

$$\begin{aligned} 0 &= -2X^T y + 2X^T Xb \\ 2X^T y &= 2X^T Xb \end{aligned} \tag{38}$$

$$b = (X^T X)^{-1} X^T y \tag{39}$$

By solving equation-39 all coefficient are obtained. While modelling of IDT, it was assumed that IDT depends on all the parameters . But significance of each independent variable on dependent variable varies and has to be verified by hypothesis testing.

6.1. Hypothesis testing

Null Hypothesis (H_0): There is no significant relationship between independent variable and dependent variable or in another words, by adding an independent variable to model will not casue any significant improvement on the dependent variable. Means $b_1 = b_2 = b_3 = ... = b_n = 0$.

Hypothesis testing done using t-test by taking confidence interval of 95%. To obtain coefficient and other other statical parameter, statmodel [46] library package is used. In which, t-values is obtained by taking ratio of coefficient and standard error. P-values (probability values) are obtained from t-values using two-tail test.

If obtained p-values is less than $\alpha = 0.05$ (5%), which means value lies in the confidence interval and obtained result or independent parameters is statically significant to dependent parameter. In 95% confidence interval or $p < 0.05$ generally null hypothesis is rejected. To select statistically significant independent variable backward elimination method is used [47]. Independent variable X_i were removed till removal does not cause drastic decrease in R^2 value.

Algorithm 2 Backward elimination

1: Use all independent variable to obtain correlation
2: find coefficient and p values of independent variable
3: **while** for any X_i : $p > 0.05$ **do**
4: Remove X_i from dataset
5: Obtain coefficient and p values

After eliminating all non-significant independent variable, obtained final correlation and result are discussed further.

7. Result and Discussion:

Most important part this study is K-means clustering. Location of cluster centroid is important for unknown fuel to select the cluster category. Movement of cluster centroids with variation in data point are mentioned in table-6. From the table it is observed that cluster centroids are very sensitive to the number of data points and fuels. But as number of data points, variety fuel, physical conditions increases cluster centroid converges to specific location and movement in centroid become less. For individual fuel, cluster centroid varies significantly but as combination of fuel increases it converges to location as in case of all fuel.

Fuels vs Axis		Temeprature	Pressure	C_P - C_H	C_P - C_S	C_S - C_S	C_S - C_H	Fuel(%)	Oxygen(%)
All	Centroid-0	1.57837308	1.51675986	1.24959195	0.41653065	0.26505381	0.94663827	-0.23781285	1.84398964
	Centroid-1	1.68012488	0.93518365	1.12307542	0.37435847	1.22535434	2.82506715	-2.88900159	-0.19240077
	Centroid-2	1.38326654	2.36889142	1.51647253	0.50549084	2.11119433	4.7278795	-0.6030923	2.56675347
Propane, Dodecane, Hexane, Butane, Heptane, Decane	Centroid-0	1.57605773	1.57148828	1.25323871	0.41774624	0.23445554	0.88665731	-0.16905458	1.88849474
	Centroid-1	1.6944866	0.91018368	1.1061271	0.36870903	1.14345847	2.65562598	-2.92404759	-0.29666696
	Centroid-2	1.35803663	2.42946412	1.55515259	0.5183842	2.16787013	4.85412445	-0.37914573	2.88516679
Propane, Dodecane, Heptane	Centroid-0	1.5444353	1.96307245	1.29529591	0.4317653	0.25076071	0.93328672	-0.11458897	1.97025854
	Centroid-1	1.35250581	2.35543233	1.56435254	0.52145085	2.26644565	5.05434215	-0.41749343	2.92323109
	Centroid-2	1.67874594	1.13744756	1.12470066	0.37490022	1.1954392	2.76577862	-2.95741867	-0.28904548
Propane	Centroid-0	1.59652546	1.25275661	1.22966736	0.40988912	0.	0.40988912	0.02766885	1.87766756]
	Centroid-1	1.42465876	3.62448678	1.4507666	0.48358887	0.	0.48358887	-0.127854	1.92640905
	Centroid-2	1.75337899	0.77994782	1.0400941	0.34669803	0.	0.34669803	1.75337899	0.77994782
Heptane	Centroid-0	1.6708165	1.22938158	1.1350263	0.3783421	0.7566842	1.89171049	-3.18913949	-0.65261478
	Centroid-1	1.65276957	0.5790902	1.15044562	0.38348187	0.76696375	1.91740937	-0.57119456	1.83443768
	Centroid-2	1.33539693	3.39234322	1.59378677	0.53126226	1.06252451	2.65631128	0.62807518	3.02529108
Dodecane	Centroid-0	1.49689856	2.6228982	1.34469596	0.44823199	2.01704393	4.48231985	-2.529175	1.33872879
	Centroid-1	1.3509045	2.26022207	1.56610377	0.52203459	2.34915565	5.22034588	-0.29324417	3.04132966
	Centroid-2	1.70214704	0.80548421	1.0960673	0.36535577	1.64410095	3.65355766	-3.07720187	-0.2855489

Table 6: Location of 3- centroids obtained using K-Means for various of fuel/fuel-combinations

From each sub-dataset 80% data is used as training set and remaining 20% is used as testing set for performing multiple regression and hypothesis testing. Obtained coefficient after complete procedure-26 is mentioned in table-7. For multiple regression greedy binary tree algorithm is implemented, in which if minimum R^2=0.85 training accuracy is not obtained then algorithm will start dividing the data into two parts based on mean temperature value of dataset. All sub-dataset satisfies that cut-off criteria and obtained coefficient are mentioned in table-7.

	Intercept	Temeprature	Pressure	C_P - C_H	C_P - C_S	C_S - C_S	C_S - C_H	Fuel(%)	Oxygen(%)	Training R^2	Test R^2
Cluster -0	37.1733	-16.1974	-0.5486	-0.9622	-0.3207	0.119	0	0.9398	-1.6978	0.95186	0.95405
	33.6319	-14.8109	-0.5521	0	0	0.1346	0	0.9329	-1.6929	0.95336	0.95254
Cluster -1	38.5577	-17.2162	-0.4416	-1.7754	-0.5918	0.2792	0	0.67	-1.1244	0.96630	0.97684
	32.4963	-14.9362	-0.4456	0	0	0.2809	0	0.6651	-1.1229	0.97353	0.96695
Cluster -2	101.8355	-42.4864	-1.0402	-19.2446	-6.4149	2.5527	-1.3095	-0.1454	-0.3402	0.88752	0.91019
	102.0124	-42.5188	-1.0379	-19.3288	-6.4429	2.5669	-1.3091	-0.1297	-0.348	0.91015	0.88578

Table 7: coefficient and accuracy obtained for 3 sub-dataset using training set of all available fuel data points

Due to generation of data points from uncertainty and sampling from generated points, causes variation in result. Such variation is acceptable as obtained accuracy for both test and train set falls within acceptable range. To observe the variation more precisely and find out the effect of sampling on coefficient, results were obtained by running the simulation 4400 times. Variation in the result is plotted by histograms:12-22. In the results, distribution of coefficients follows the nearly Gaussian distribution. for coefficient distribution deviates due to classification error. But as overall accuracy falls within acceptable range such results are considerable.

Sub-dataset - 0 Sub-dataset - 1 Sub-dataset - 2

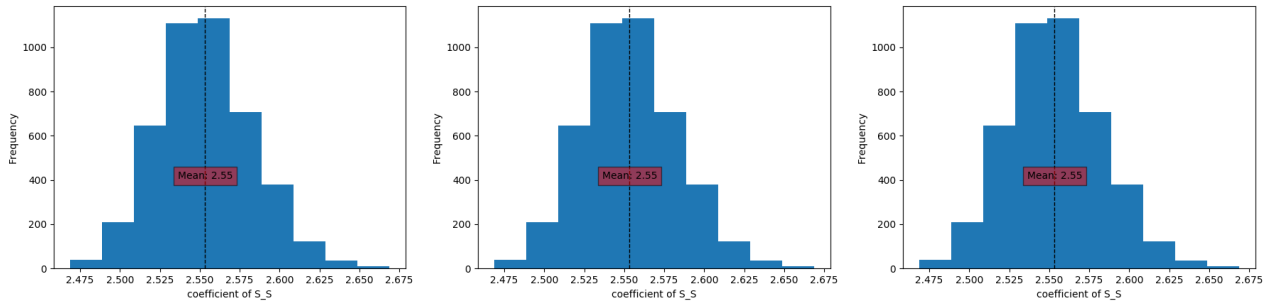


Figure 12: Secondary-Secondary Carbon bond coefficient of sub-dataset

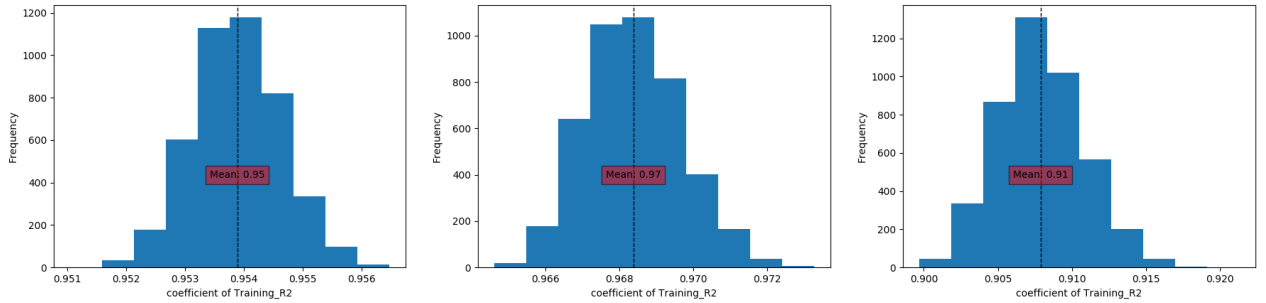


Figure 13: Training accuracy of sub-dataset

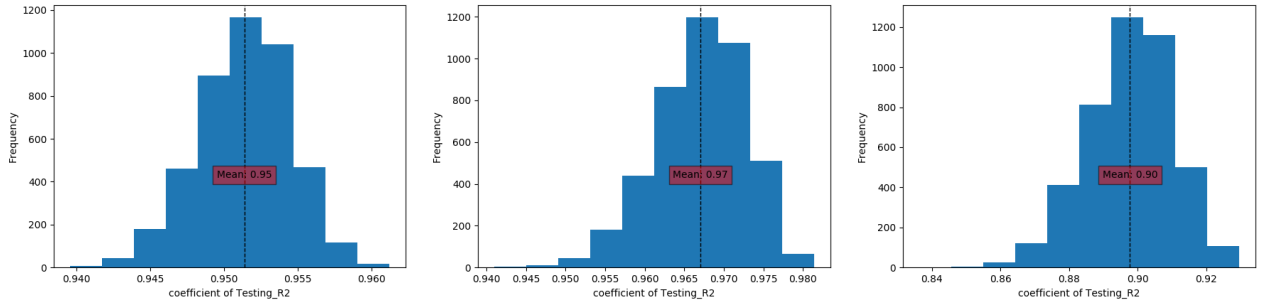


Figure 14: Testing accuracy of sub-dataset

Sub-dataset - 0 Sub-dataset - 1 Sub-dataset - 2

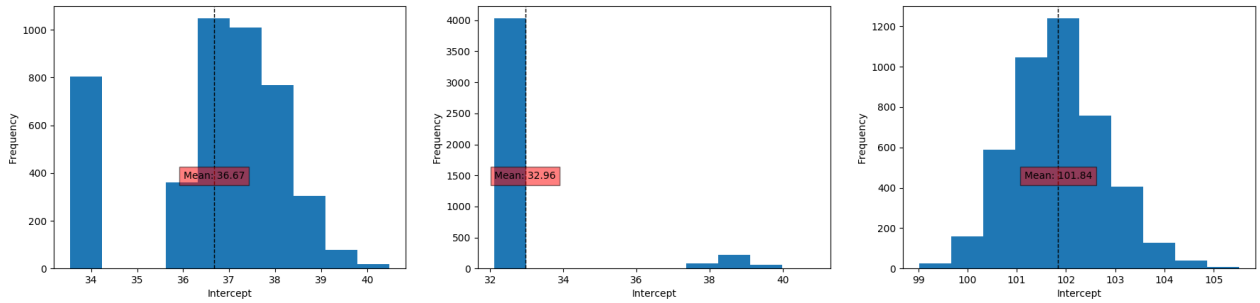


Figure 15: Intercept of sub-dataset

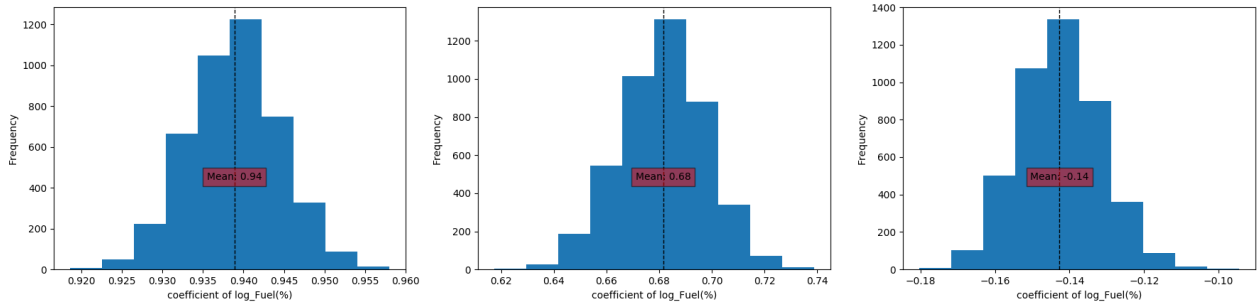


Figure 16: Fuel mole fraction term coefficient of sub-dataset

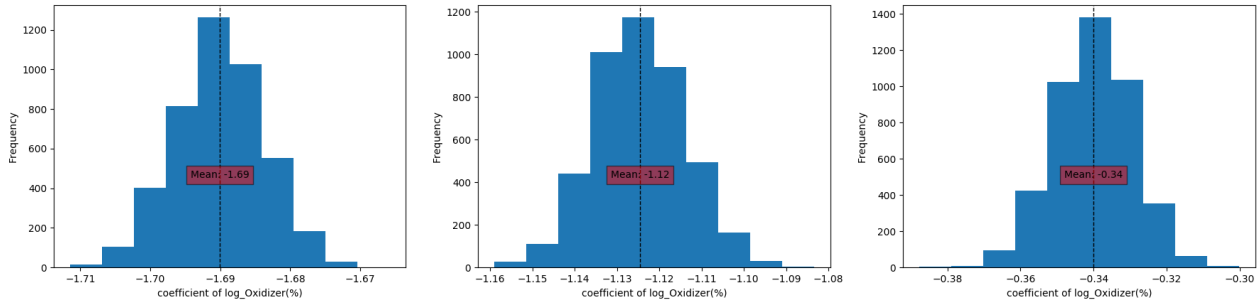


Figure 17: Oxygen term coefficient of sub-dataset

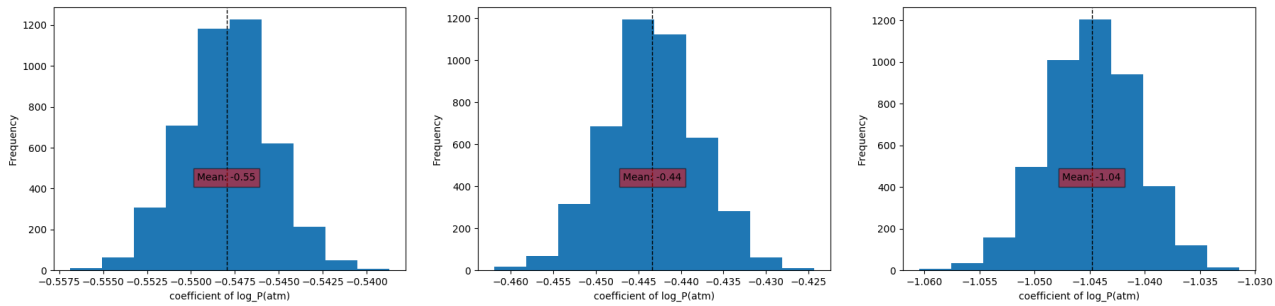


Figure 18: Pressure term coefficient of sub-dataset

Sub-dataset - 0 Sub-dataset - 1 Sub-dataset - 2

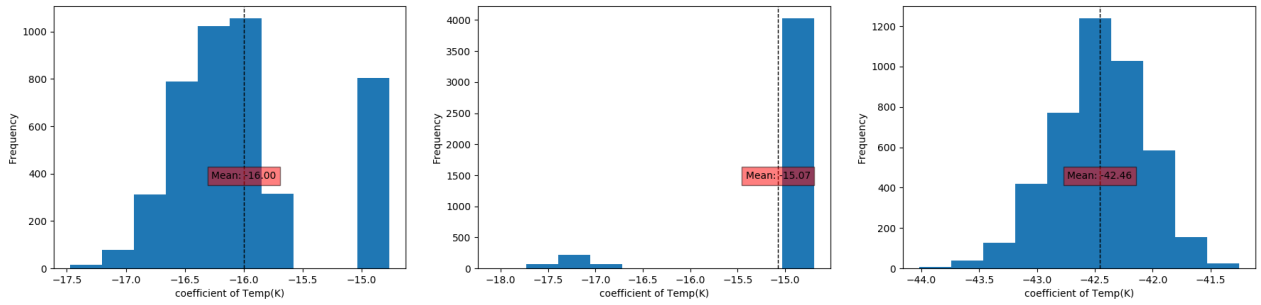


Figure 19: Temperature term coefficient of sub-dataset

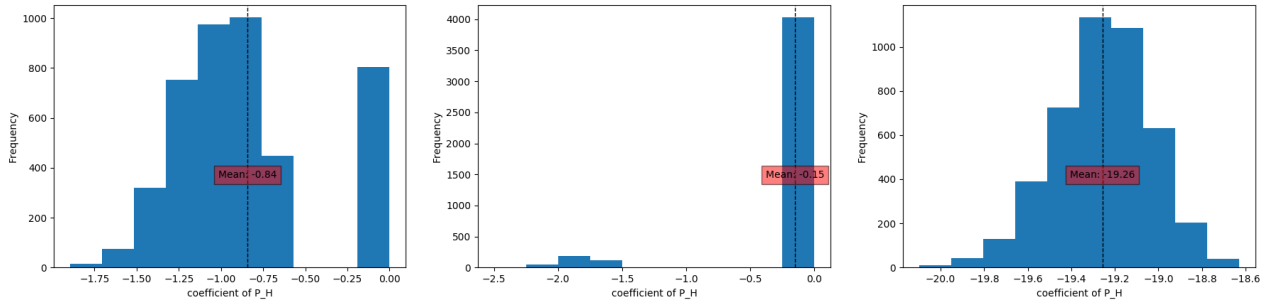


Figure 20: Primary Carbon and Hydrogen bond coefficient of sub-dataset

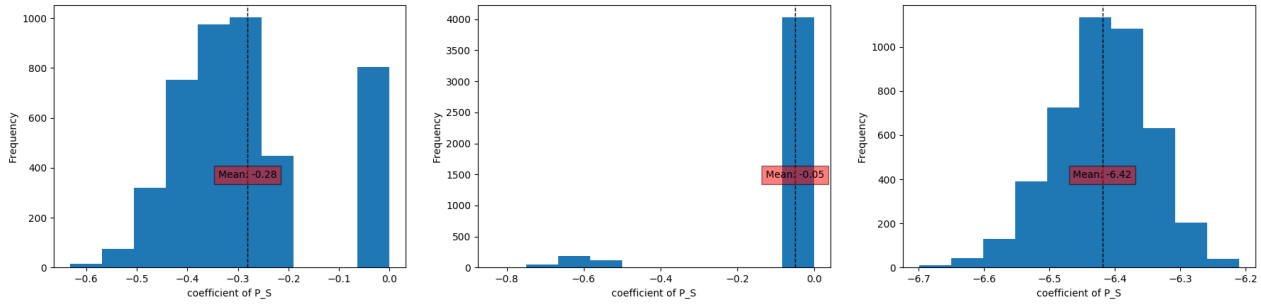


Figure 21: Primary-Secondary Carbon bond coefficient of sub-dataset

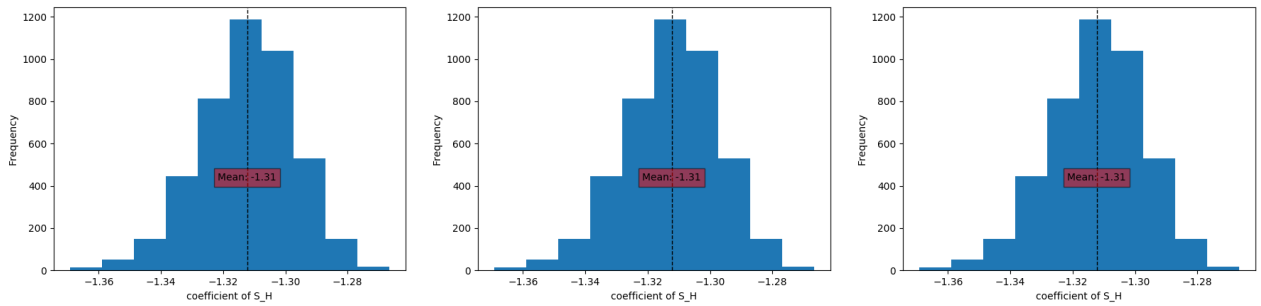


Figure 22: Secondary Carbon and Hydrogen bond coefficient of sub-dataset

Formulation of Ignition delay time is acquired using mean value of such distribution. After 4200 repetitive result mean value of coefficient converges to specific value as given in table-8. IDT correlation obtained from converged coefficient value- 8 is given as below. Associated value of centroids are mentioned in table-6 for 'all fuel' case:

- For centroid-0:

$$\tau = e^{36.68} \cdot \left(\frac{T}{T_0}\right)^{-15.99} \cdot \left(\frac{P}{P_0}\right)^{-0.55} \cdot X_{Fuel}^{0.94} \cdot X_{O_2}^{-1.69} \cdot e^{(C_P H)^{-0.84}} \cdot e^{(C_P C_S)^{-0.28}} \cdot e^{(C_S C_S)^{-0.13}} \quad (40)$$

- For centroid-1:

$$\tau = e^{32.96} \cdot \left(\frac{T}{T_0}\right)^{-15.08} \cdot \left(\frac{P}{P_0}\right)^{-0.44} \cdot X_{Fuel}^{0.68} \cdot X_{O_2}^{-1.12} \cdot e^{(C_P H)^{-0.16}} \cdot e^{(C_P C_S)^{-0.05}} \cdot e^{(C_S C_S)^{-0.29}} \quad (41)$$

- For centroid-2:

$$\tau = e^{101.85} \cdot \left(\frac{T}{T_0}\right)^{-42.46} \cdot \left(\frac{P}{P_0}\right)^{-1.04} \cdot X_{Fuel}^{-0.14} \cdot X_{O_2}^{-0.34} \cdot e^{(C_P H)^{-19.26}} \cdot e^{(C_P C_S)^{-6.42}} \cdot e^{(C_S C_S)^{-2.55}} \cdot e^{(C_S H)^{-1.31}} \quad (42)$$

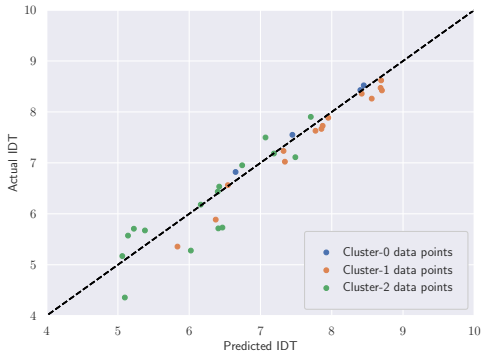
	Number of Simulation	Constant	P_H	P_S	S_H	S_S	Temp(K)	log_Fuel(%)	log_Oxidizer(%)	log_P(atm)	Testing_R2	Training_R2
Cluster-0	500	36.67	-0.84	-0.28	0	0.13	-16	0.94	-1.69	-0.55	0.95	0.95
	1000	36.72	-0.86	-0.29	0	0.13	-16.02	0.94	-1.69	-0.55	0.95	0.95
	1500	36.67	-0.84	-0.28	0	0.13	-15.99	0.94	-1.69	-0.55	0.95	0.95
	2000	36.66	-0.84	-0.28	0	0.13	-15.99	0.94	-1.69	-0.55	0.95	0.95
	2500	36.65	-0.84	-0.28	0	0.13	-15.99	0.94	-1.69	-0.55	0.95	0.95
	3000	36.65	-0.84	-0.28	0	0.13	-15.99	0.94	-1.69	-0.55	0.95	0.95
	3500	36.67	-0.84	-0.28	0	0.13	-16	0.94	-1.69	-0.55	0.95	0.95
	4000	36.68	-0.85	-0.28	0	0.13	-15.99	0.94	-1.69	-0.55	0.95	0.95
	4200	36.68	-0.84	-0.28	0	0.13	-15.99	0.94	-1.69	-0.55	0.95	0.95
	4395	36.68	-0.84	-0.28	0	0.13	-15.99	0.94	-1.69	-0.55	0.95	0.95
Cluster-1	500	33.16	-0.21	-0.07	0	0.28	-15.15	0.68	-1.12	-0.44	0.97	0.97
	1000	33.04	-0.18	-0.05	0	0.29	-15.1	0.68	-1.12	-0.44	0.97	0.97
	1500	33.01	-0.17	-0.06	0	0.29	-15.09	0.68	-1.12	-0.44	0.97	0.97
	2000	32.99	-0.16	-0.05	0	0.29	-15.08	0.68	-1.12	-0.44	0.97	0.97
	2500	32.98	-0.16	-0.05	0	0.29	-15.08	0.68	-1.12	-0.44	0.97	0.97
	3000	32.97	-0.16	-0.05	0	0.29	-15.08	0.68	-1.12	-0.44	0.97	0.97
	3500	32.97	-0.16	-0.05	0	0.29	-15.08	0.68	-1.12	-0.44	0.97	0.97
	4000	32.96	-0.16	-0.05	0	0.29	-15.08	0.68	-1.12	-0.44	0.97	0.97
	4200	32.96	-0.16	-0.05	0	0.29	-15.08	0.68	-1.12	-0.44	0.97	0.97
	4395	32.96	-0.16	-0.05	0	0.29	-15.08	0.68	-1.12	-0.44	0.97	0.97
Cluster-2	500	101.83	-19.26	-6.42	-1.31	2.55	-42.46	-0.14	-0.34	-1.04	0.9	0.91
	1000	101.85	-19.26	-6.42	-1.31	2.55	-42.46	-0.14	-0.34	-1.04	0.9	0.91
	1500	101.86	-19.26	-6.42	-1.31	2.55	-42.46	-0.14	-0.34	-1.04	0.9	0.91
	2000	101.84	-19.26	-6.42	-1.31	2.55	-42.46	-0.14	-0.34	-1.04	0.9	0.91
	2500	101.84	-19.26	-6.42	-1.31	2.55	-42.46	-0.14	-0.34	-1.04	0.9	0.91
	3000	101.86	-19.26	-6.42	-1.31	2.55	-42.46	-0.14	-0.34	-1.04	0.9	0.91
	3500	101.85	-19.26	-6.42	-1.31	2.55	-42.46	-0.14	-0.34	-1.04	0.9	0.91
	4000	101.84	-19.26	-6.42	-1.31	2.55	-42.46	-0.14	-0.34	-1.04	0.9	0.91
	4200	101.84	-19.26	-6.42	-1.31	2.55	-42.46	-0.14	-0.34	-1.04	0.9	0.91
	4395	101.85	-19.26	-6.42	-1.31	2.55	-42.46	-0.14	-0.34	-1.04	0.9	0.91

Table 8: Coefficient values obtained after given number of simulations for different sub-dataset which shows convergence of coefficient values

For verification of correlation and frame work, smaller set dataset of heptane and hexadecane from, main dataset used to verify the result. All fuel data points used for training the model heptane and hexadecane are also part of it. After completer procedure mentioned as in 26, observed result shows excellent match with predicted value. For cluster-0,1 relative error is bounded between

0.8%-3% and 0.4%-5% respectively 23. For cluster-2 error varied from 0.09% to 16% which is high as obtained coefficient of determination is also low.

(a) IDT prediction of hexane and hexadecane fuel by Framework using all fuel data



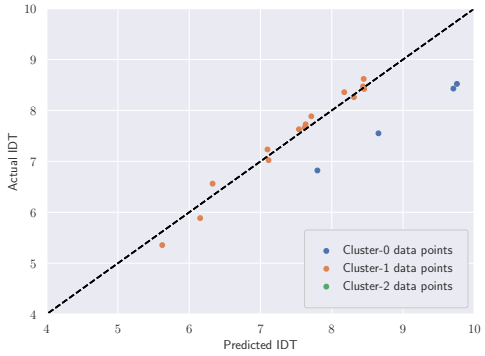
(b) IDT prediction of hexane and hexadecane fuel using correlation using all fuel data

Figure 23: Prediction of IDT of hexane and hexadecane using correlation and Framework

Result of IDT from correlation shows slightly more deviation. But it follows the same trend as of framework result. Good prediction of hexane and hexadecane is expected as correlation already contains those fuel detail. The goal is present study is to predict IDT for new unknown fuel.

To check prediction of unknown straight chain alkane fuel, Hexadecane and Hexane is used as unknown fuel which would not be part of training and testing set. Model is trained using remaining all fuel. To attain main objective and check behaviour of framework as well as correlational result are mentioned in figure-24.

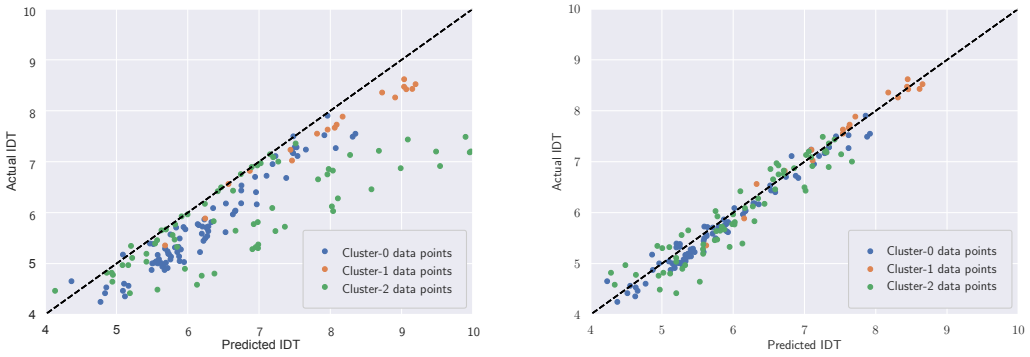
(a) IDT prediction of hexane and hexadecane fuel by Framework using fuel other than hexane and hexadecane



(b) IDT prediction of hexane and hexadecane fuel using correlation using fuel other than hexane and hexadecane.

Figure 24: Prediction of IDT of hexane and hexadecane using correlation and Framework

Result of figure-24 shows that prediction using framework again gives excellent fit but result obtained using correlation incurred considerable error in prediction. Error in correlation result might be caused due to the movement of centroids which shows that Accuracy of correlational depends also on centroid value.



(a) IDT prediction of all fuel other than propane, heptane and dodeacne by Framework using propane, heptane and dodeacne as learning set

(b) IDT prediction of all fuel other than propane, heptane and dodeacne using correlation using propane, heptane and dodeacne as learning set.

Figure 25: Prediction of IDT of hexane and hexadecane using correlation and Framework

To verify extreme case, propane, heptane and dodecane were used to train the model as it almost covers wide range of condition and different length of alkanes. Rest of all fuel were used as unknown fuel to test the prediction. Observed result was quite unexpected, which is shown in fig-25. Obtained result from framework is quite scattered specifically for cluster-2. Whereas result obtained using correlation 40,41,42 shows excellent agreement of predicted result with experimental result. Centroids were obtained from propane, heptane and dodecane. Such pair of centroid and correlation is important for prediction IDT of new fuel.

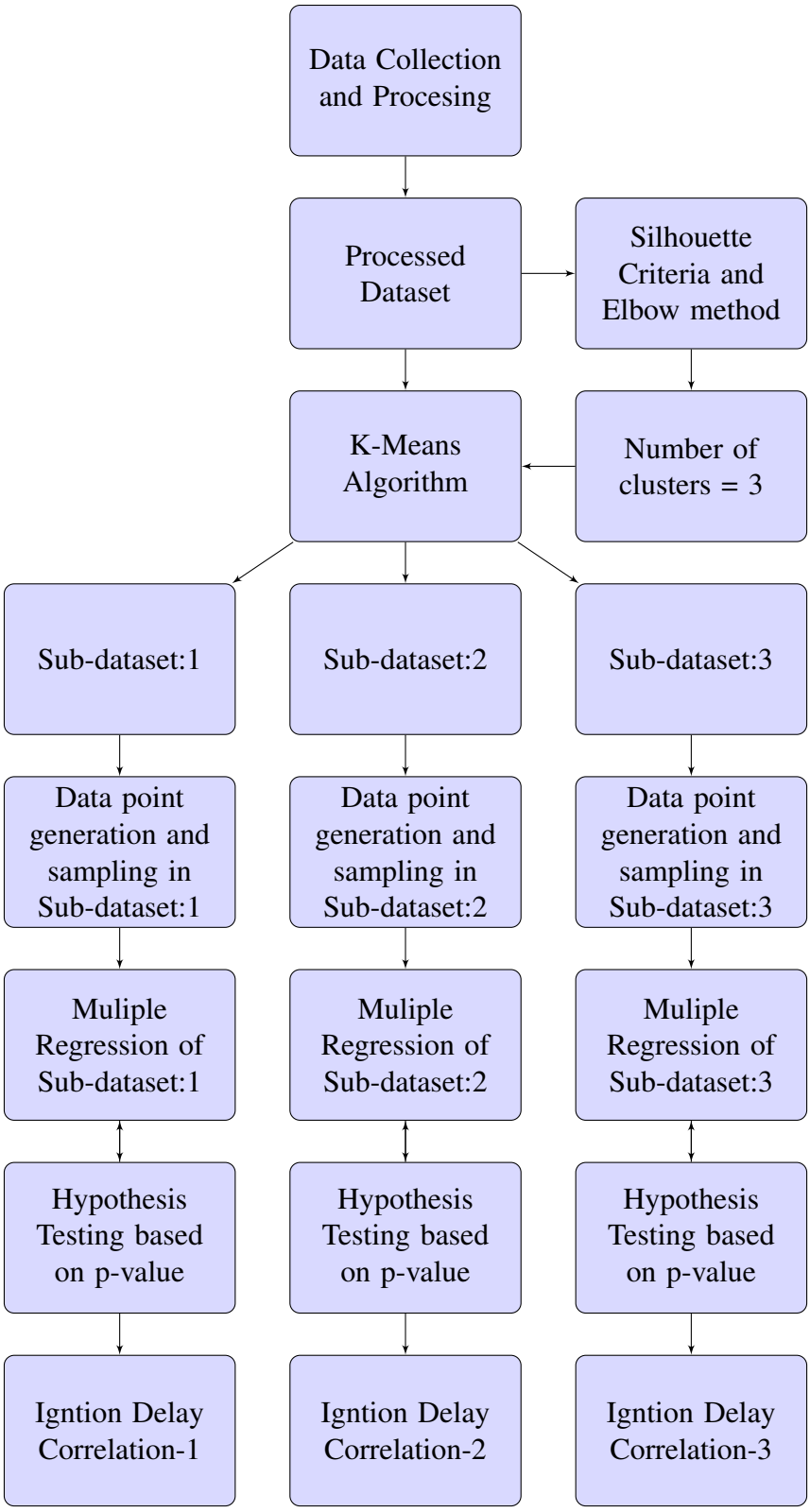


Figure 26: Flowchart of Ignition Delay Time Prediction Framework

References:

[1] T. Lu and C. K. Law, "Toward accommodating realistic fuel chemistry in large-scale computations," *Progress in Energy and Combustion Science*, vol. 35, no. 2, pp. 192–215, 2009.

[2] D. C. Horning, "A study of the high-temperature auto-ignition and thermal decomposition of hydrocarbons," vol. Report No. TSD-135, 2001.

[3] F. Khaled and A. Farooq, "On the universality of ignition delay times of distillate fuels at high temperatures: A statistical approach," *Combustion and Flame*, vol. 210, pp. 145–158, 2019.

[4] S. S. Goldsborough, "A chemical kinetically based ignition delay correlation for iso-octane covering a wide range of conditions including the ntc region,"

[5] Z. Zhao, Z. Chen, and S. Chen, "Correlations for the ignition delay times of hydrogen/air mixtures," *Chinese science bulletin*, vol. 56, no. 2, pp. 215–221, 2011.

[6] W. Ji, P. Zhao, T. He, X. He, A. Farooq, and C. K. Law, "On the controlling mechanism of the upper turnover states in the ntc regime," *Combustion and Flame*, vol. 164, pp. 294–302, 2016.

[7] N. Shah, P. Zhao, D. DelVescovo, and H. Ge, "Prediction of autoignition and flame properties for multicomponent fuels using machine learning techniques," tech. rep., SAE Technical Paper, 2019.

[8] K. Dussan, S. H. Won, A. D. Ure, F. L. Dryer, and S. Dooley, "Chemical functional group descriptor for ignition propensity of large hydrocarbon liquid fuels," *Proceedings of the Combustion Institute*, vol. 37, no. 4, pp. 5083–5093, 2019.

[9] S. Chinta, A. Sivaram, and R. Rengaswamy, "Prediction error-based clustering approach for multiple-model learning using statistical testing," *Engineering Applications of Artificial Intelligence*, vol. 77, pp. 125–135, 2019.

[10] E. V. Anslyn and D. A. Dougherty, *Modern physical organic chemistry*. University science books, 2006.

[11] S. J. Blanksby and G. B. Ellison, "Bond dissociation energies of organic molecules," *Accounts of chemical research*, vol. 36, no. 4, pp. 255–263, 2003.

[12] "Propane mechanism - aramcomech2.0," Available at http://www.nuigalway.ie/media/researchcentres/combustionchemistrycentre/files/mechanismdownloads/aramcomech2/AramcoMech2.0.mech.

[13] "Butane mechanism - aramcomech1.3," Available at http://www.nuigalway.ie/media/researchcentres/combustionchemistrycentre/files/mechanismdownloads/c4_49.dat.

[14] "Pentane mechanism - aramcomech1.3," Available at http://www.nuigalway.ie/media/researchcentres/combustionchemistrycentre/files/mechanismdownloads/nuigmech_c5_july2015_1215.chem.

[15] K. Zhang, C. Banyon, C. Togbé, P. Dagaut, J. Bugler, and H. J. Curran, "Hexane mechanism," Available at https://ars.els-cdn.com/content/image/1-s2.0-S0010218015002576-mmc1.zip.

[16] K. Zhang, C. Banyon, J. Bugler, H. J. Curran, A. Rodriguez, O. Herbinet, F. Battin-Leclerc, C. B'Chir, and K. A. Heufer, "Heptane mechanism," Available at https://ars.els-cdn.com/content/image/1-s2.0-S0010218016301560-mmc1.zip.

[17] W. J. P. O. H. H. J. C. Westbrook, C. K. and E. J. Silke, "Octane to hexa-decane mechanisms," Available at https://combustion.llnl.gov/archived-mechanisms/alkanes/c8c16-nalkanes.

[18] D. C. Horning, D. Davidson, and R. Hanson, *A study of the high-temperature autoignition and thermal decomposition of hydrocarbons*. PhD thesis, Stanford University Stanford, California, 2001.

[19] D. C. Horning, D. Davidson, and R. Hanson, "Study of the high-temperature autoignition of n-alkane/o/ar mixtures," *Journal of Propulsion and Power*, vol. 18, no. 2, pp. 363–371, 2002.

[20] D. Davidson, J. Herbon, D. Horning, and R. Hanson, "Oh concentration time histories in n-alkane oxidation," *International journal of chemical kinetics*, vol. 33, no. 12, pp. 775–783, 2001.

[21] K. Y. Lam, *Shock tube measurements of oxygenated fuel combustion using laser absorption spectroscopy*. PhD thesis, Stanford University, 2013.

[22] K.-Y. Lam, Z. Hong, D. Davidson, and R. Hanson, "Shock tube ignition delay time measurements in propane/o2/argon mixtures at near-constant-volume conditions," *Proceedings of the Combustion Institute*, vol. 33, no. 1, pp. 251–258, 2011.

[23] S. M. Burke, U. Burke, R. Mc Donagh, O. Mathieu, I. Osorio, C. Keesee, A. Morones, E. L. Petersen, W. Wang, T. A. DeVerter, *et al.*, "An experimental and modeling study of propene oxidation. part 2: Ignition delay time and flame speed measurements," *Combustion and Flame*, vol. 162, no. 2, pp. 296–314, 2015.

[24] D. Davidson, S. Ranganath, K.-Y. Lam, M. Liaw, and Z. Hong, "Ignition delay time measurements of normal alkanes and simple oxygenates," *Journal of Propulsion and Power*, vol. 26, no. 2, pp. 280–287, 2010.

[25] B. Gauthier, D. F. Davidson, and R. K. Hanson, "Shock tube determination of ignition delay times in full-blend and surrogate fuel mixtures," *Combustion and Flame*, vol. 139, no. 4, pp. 300–311, 2004.

[26] D. Davidson, M. Oehlschlaeger, and R. Hanson, "Methyl concentration time-histories during iso-octane and n-heptane oxidation and pyrolysis," *Proceedings of the Combustion Institute*, vol. 31, no. 1, pp. 321–328, 2007.

[27] S. S. Vasu, D. F. Davidson, and R. K. Hanson, "Oh time-histories during oxidation of n-heptane and methylcyclohexane at high pressures and temperatures," *Combustion and Flame*, vol. 156, no. 4, pp. 736–749, 2009.

[28] D. Davidson, Z. Hong, G. Pilla, A. Farooq, R. Cook, and R. Hanson, "Multi-species time-history measurements during n-heptane oxidation behind reflected shock waves," *Combustion and flame*, vol. 157, no. 10, pp. 1899–1905, 2010.

[29] D. R. Haylett, *The development and application of aerosol shock tube methods for the study of low-vapor-pressure fuels*. Stanford University, 2011.

[30] D. Davidson, D. Haylett, and R. Hanson, "Development of an aerosol shock tube for kinetic studies of low-vapor-pressure fuels," *Combustion and Flame*, vol. 155, no. 1-2, pp. 108–117, 2008.

[31] D. R. Haylett, D. F. Davidson, and R. K. Hanson, "Ignition delay times of low-vapor-pressure fuels measured using an aerosol shock tube," *Combustion and Flame*, vol. 159, no. 2, pp. 552–561, 2012.

[32] D. Haylett, D. Davidson, and R. Hanson, "Second-generation aerosol shock tube: an improved design," *Shock Waves*, vol. 22, no. 6, pp. 483–493, 2012.

[33] D. Haylett, D. Davidson, and R. Hanson, "A second-generation aerosol shock tube for combustion research," in *48th AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition*, p. 196, 2010.

[34] D. Jackson, D. Davidson, and R. Hanson, "Application of an aerosol shock tube for the kinetic studies of n-dodecane/nano-aluminum slurries," in *44th AIAA/ASME/SAE/ASEE joint propulsion conference & exhibit*, p. 4767, 2008.

[35] S. S. Vasu, D. F. Davidson, Z. Hong, V. Vasudevan, and R. K. Hanson, "n-dodecane oxidation at high-pressures: Measurements of ignition delay times and oh concentration time-histories," *Proceedings of the Combustion Institute*, vol. 32, no. 1, pp. 173–180, 2009.

[36] S. S. Vasu, *Measurements of ignition times, OH time-histories, and reaction rates in jet fuel and surrogate oxidation systems*. PhD thesis, PhD Thesis, Stanford University, California, United States, 2010.

[37] D. Haylett, D. Davidson, R. Cook, Z. Hong, W. Ren, S. Pyun, and R. Hanson, "Multi-species time-history measurements during n-hexadecane oxidation behind reflected shock waves," *Proceedings of the Combustion Institute*, vol. 34, no. 1, pp. 369–376, 2013.

[38] D. Haylett, R. Cook, D. Davidson, and R. Hanson, "Oh and c2h4 species time-histories during hexadecane and diesel ignition behind reflected shock waves," *Proceedings of the Combustion Institute*, vol. 33, no. 1, pp. 167–173, 2011.

[39] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of chemical information and computer sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[40] G. Landrum, "Rdkit: Open-source cheminformatics,"

[41] G. Ogbuabor and F. Ugwoke, "Clustering algorithm for a healthcare dataset using silhouette score value," *International Journal of Computer Science & Information Technology*, vol. 10, no. 2, pp. 27–37, 2018.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[43] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[44] R. C. de Amorim and C. Hennig, "Recovering the number of clusters in data sets with noise features using feature rescaling factors," *Information Sciences*, vol. 324, pp. 126–145, 2015.

[45] N. V. Chawla, "Data mining for imbalanced datasets: An overview," in *Data mining and knowledge discovery handbook*, pp. 875–886, Springer, 2009.

[46] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

[47] J. H. McDonald, *Handbook of biological statistics*, vol. 2. 2009.