Understanding Atmospheric Chemistry using Graph-Theory, Visualisation and Machine Learning.

Dan Ellis

March 2020

Veritatem inquirenti, semel in vita de omnibus, quantum fieri potest, esse dubitandu:
In order to seek truth, it is necessary once in the course of our life, to doubt, as far as possible, of all things.

- Descartes, Rene, *Principles of Philosophy*

Contents

1 Introduction **1**

1.1 Background . . . 4

1.1.1 A Preface on Humanity and the Climate . . . 4

1.1.2 Formation of the Atmosphere . . . 4

1.1.3 Rise of the Homo Spiens ('Wise Man') . . . 4

1.2 Motivation (How the atmosphere affects us) . . . 5

1.3 Air Quality - it is the air we breathe . . . 5

1.4 Protection - Ozone and its role . . . 6

1.4.1 The NOx cycle . . . 6

1.5 Changing Climate . . . 7

1.5.1 HOx Cycle . . . 7

1.6 Modelling the Earth . . . 7

1.6.1 Earth System Models (ESM) . . . 8

1.6.2 The box model. . . . 9

1.6.2.1 Chemical Mechanisms . . . 9

1.6.3 Numerical integration . . . 10

1.6.3.1 Non-Stiff Equations . . . 10

1.6.3.2 Numerically stiff equations (atmospheric chemistry) . . . 10

1.6.4 The model development cycle . . . 10

1.6.5 The Dynamically Simple Model of Atmospheric Chemical Complexity . . . 11

1.7 Thesis Layout . . . 11

Appendices **13**

A Supplementary Mathematics **15**

A.1 PCA . . . 15

A.1.1 Statistics . . . 15

A.1.2 Matrices and Eigenvectors . . . 16

A.2 t-SNE . . . 17

A.2.1 Student T distribution . . . 17

A.2.1.1 T-Score . . . 17

A.2.2 Kullback-Leiber (KL) divergence . . . 17

B Neural Network Activation Functions **19**

B.1 Binary Step 19

B.2 Linear 20

B.3 Sigmoid / Logistic 20

B.4 Hyperbolic Tangent 21

B.5 Rectified Linear Unit 21

B.6 Swish 22

B.7 **A note on backpropagation** 22

C Miscellaneous **23**

C.1 Correspondance with Mike Jenkin 23

C.2 Functional Groups 27

D Chapter Keywords **29**

D.1 Introduction 29

D.2 Applying Visual Analytics to the Atmospheric Chemistry Network 30

D.3 Computational Learning, Visualisation and Clustering: 30

Chapter 1

Introduction

"In the beginning the Universe was created. This has made a lot of people very angry and been widely regarded as a bad move"

- Douglas Adams, *The Restaurant at the End of the Universe*

1.1 Background

1.1.1 A Preface on Humanity and the Climate

The development of humanity is not unlike the chirography of an Aristotelian tragedy. It starts with a simple/primitive species cradling a noble cause - to improve their chances of survival. Here the protagonist (humankind) develops a fatal flaw: an insecurity and latent distruction of their home due to a sudden rise to power. Having acknowleged this flaw, we now strive to imporve our understanding of the universe, correct past mistakes and stem the tide of inevitable change.

With tragedy being an imitation not of humanity, but of action and life, happiness and misery, it is only expected that such a comparison to our current affairs should stir feelings of catharsis when exploring our need for research and scientific advancement. It is with that I begin this thesis with the begining of the planet, its atmosphere and consequently the start of humankind.

1.1.2 Formation of the Atmosphere

4.5 billion years ago the Earth began as a disk of dust and gas orbiting our sun. The movement of such gasses produces a resonant drag instability, which causes them to clump together [Hopkins and Squire, 2018; Woo, 2018]. As these 'clumps' become denser, other forces come in to play and further increase their size. These eventually produced the hot mix of gas and solid which was to become Earth. As the Earth cooled, the vollotile componenets of the primodial gas cloud surrounding it begin to form an atmosphere. At this point in time oxygen was not only absent in the atmopshere, but also had many sinks within the Earths anoxidised crust. It was not until oxygenic photosynthesis ([Peretó, 2011]) that the concentrations of oxygen in the atmosphere started to increase. Eventually the development of multicellular cyanobacteria[1] resulted in biologically induced oxygen accumelating in the atmosphere, [University of Zurich, 2013]. This led to the most significant climate event in the planets history: the Great Oxigenation Event (2.5 billion years ago), [Planavsky et al., 2014]. This increase of oxygen allowed oragnisms to become larger and more active, eventually resulting in the human race.

1.1.3 Rise of the Homo Spiens ('Wise Man')

About x million years ago there were many varieties of the homo genus. With the development of the human brain, energy transfer changed. A larger brain required more fuel, and therefore with

[1]The phylum of phtosynthetic prokaryotic (cells not containing a distinct nucleus) bacteria - e.g. blue-green algae

the development of cooking[2] humans were able to increase their... This form on ingenuity eventually resulted in the ... industrial revolution. This led to the first know source of indoor air pollution. Ever since we have experienced issues regarding to this...

As part of this air pollution and climate have always been a concern for the human race. Concerns about lead in the air can be documented back as far as 6000 years ago [se ref,], in ancient Rome [1145] and in 1285 where after a visit from Queen of England to a coal burning to Nottingham, the first air pollution act was deployed [1147]. air pollution = animals air qulaity pollicy kingxx With the this increased capability, a language capable of communicating information, allowing for the ability to not only hunt larger prey but also. Ability to metaphorical, allowed fruther knwoplege transfer , cvave paings and metaphoirical for people over 150 .... REFERENCES TO OTHER CHAPTERS... - vis - accounting via metaphors - and an interest in science, and atmosphere

1.2 Motivation (How the atmosphere affects us)

The atmosphere makes up an integral part of the earth system. It is responsible for shielding the earth from harmful radiation, allowing convection and weather...

1.3 Air Quality - it is the air we breathe

The atmosphere consists mainly of Nitrogen and ... However it also provides us with Oxygen, an element we rely on for the production of energy.

It therefore follows that the contamination of a persons air supply can have drastic consequences on their health. The world health organisation [Organization, 2018] state that an estimated 4.2 million premature deaths globally are linked to ambient air pollution[3] (Figure 1.1).

As low air quality can cause respiotory problems, heart disease, strokes, cancer and chronic obstructive pulmonary disease Organization [2018], it is shocking to see that over 80% of people living in urbain environmets[4] are exposed to air quality levels which exceed the reccomended limits by WHO - with low-income cities being the most impacted [**?**]. Pollutants range from particulate matter (PM) to ozone (O_3), nitrogen (NO_2) and sulpher dioxides (SO_2) - some of which will be discussed in Section 1.4.

[2] The first known case of indoor air pollution

[3] A similar number can also be attributed to indoor air pollution - which also falls under the umbrella term of Air-Quality.

[4] Which measure the levels of air pollution.

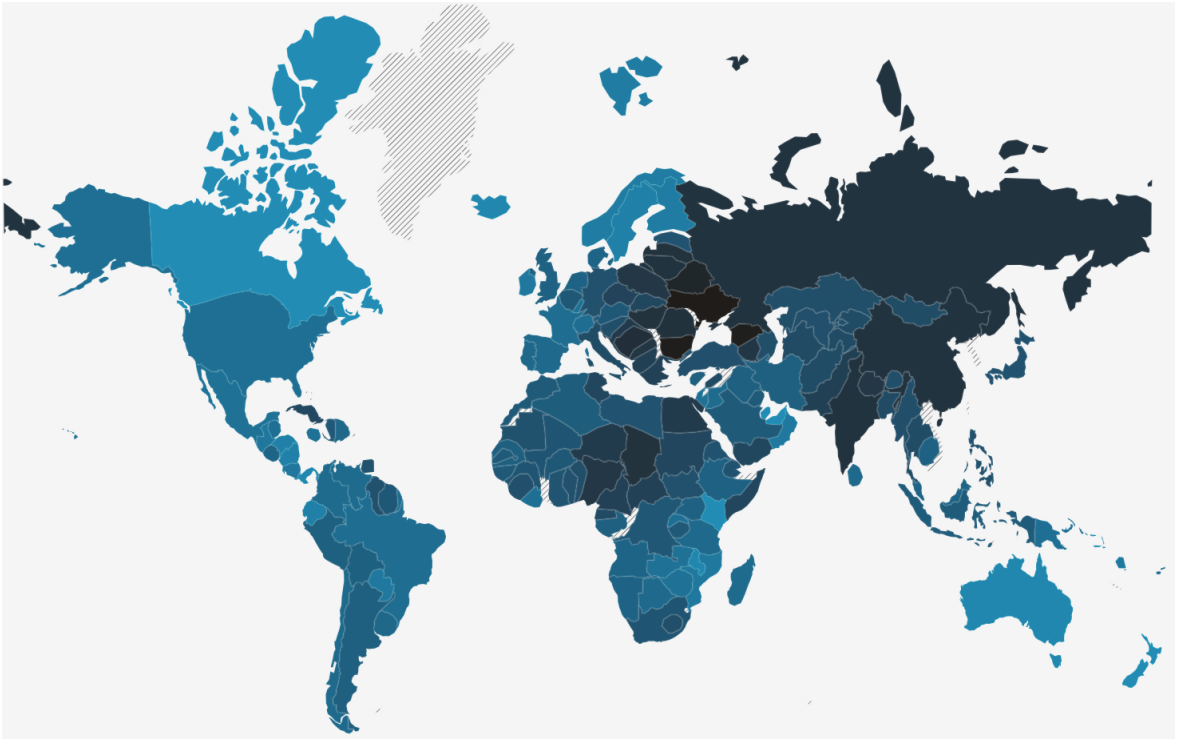


Figure 1.1: **Reported eaths attributed to air pollution by country (2016)**A cartogram cloropleth showing the number of premature deaths attributed to ambient air pollution per 100,000. The colour bar range is from 9 (light blue) to 170 (navy) people. Data Source:[**?**]

1.4 Protection - Ozone and its role

Ozone has two roles within the atmosphere. High up in the stratosphere it servers as a barrier to dangerous ultraviolet radiation. The importance of this was discovered in (HOLE PAPER) where the release of CloroFlouroCarbons from deodorants produced ... Here it is created and destroyed by the chapman cycle. This is a steady state cycle which... However within the troposhere (<15k?) the production and loss of ozone has a direct impact on human life. Polluted environments, such as industrial London, SMOG, Clean air act.

1.4.1 The NOx cycle

Nitrogen Oxides (NOx) come predominantly from motor verhicles and power stations and can cause respiotory problems in children and asmatics [se1261]. The effects of this were witnessed in early 2020 when the COV-19 corona virus disrupted travel across mainland china - Figure 1.2.

In addition to its effects on the respiratory system, notirogen oxides play a key role for Ozone formation in the troposphere. As ozone is not emitted directly, its creation depends on the chemical reacitons of other species (secondary pollutant).

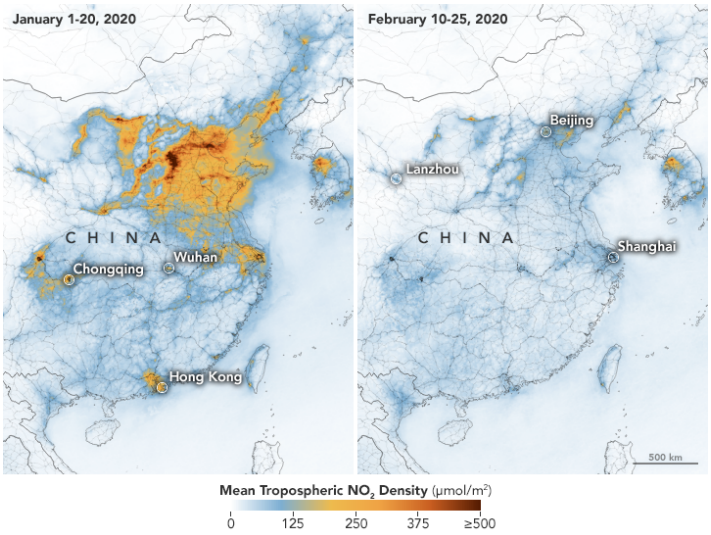


Figure 1.2: **Changes in NOx concentrations due to anthropogenic emissions.** A reduction in activity and trasport resutls in a large decrease of Nitrogen dioxide concentrations in the troposphere. Source: [Stevens, 2020]

1.5 Changing Climate

Changing climate is an <urgent> issue. Rising temperatures result in an increase of glacier melting, rising sea-levels, extreme weather and the extinction of many species, through the destruction of habitats, or creation of inhospitable conditions [refrefrefref]

[Höhne et al., 2020]

The hydroxyl radical (OH) is an important oxidising molecule that initiates the removal of pollutants from the atmosphere. OH is fairly ubiquitous in the atmosphere during the daytime, yet because it is very reactive it typically has an atmospheric lifetime of less than 1 second and an typical ambient concentration of 1 ppt [161] The reaction of molecules such as VOCs with OH forms more oxidised products, that are more water soluble, and hence it facilitates the removal of pollutants from the atmosphere by wet deposition [188].

1.5.1 HOx Cycle

1.6 Modelling the Earth

In the previous section the air quality and its detrimental effects on human health was seen to influence polcicy for cities and industry. Koyoto, Islands suing powerstations. For a policy to be passed there needs to not only evidence of the problem, but a strong suggestion that any proposed changes will have the desired effect. As it is not possible to perform experiments on complex, and often unknown,

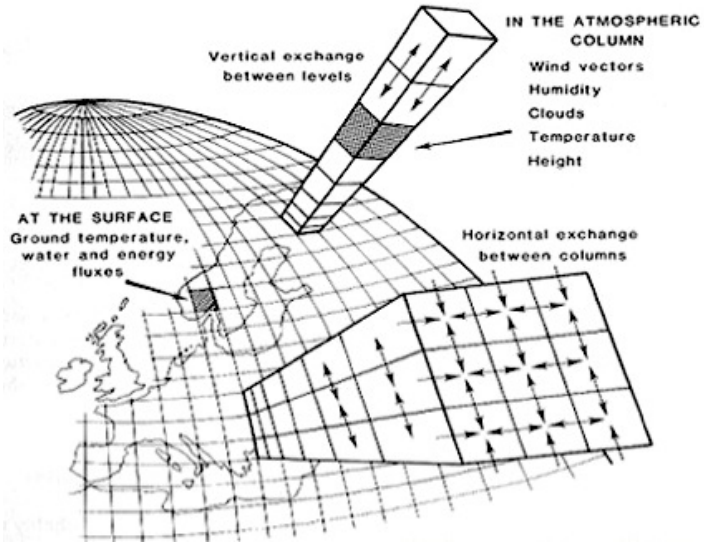


Figure 1.3: **A diagram showing the longitudal, lateral and vertical decomposition of a 3D global model.** Source: [Henderson-Sellers, 2015]

chemistry at every location on the planet, we are forced to rely on the numerical simulation of the Earth System, and the constituent parts within it.

1.6.1 Earth System Models (ESM)

ESMs are models capable of predict past or future interactions of the planetary system. They represnt our foremost understanding of the complex interplay between land-surface (geosphere), ocean (hydrosphere), ice (cryosphere) and the air (atmosphere), and act as a surrogate to manual experimentation - which is just not possible on the global scale. ESMs can be split into their individial parts. One example of this is the Chemistry section of the Goddard Earth Observing System (an integrated ESM and data assimulation model hosted by NASAs Goddard space flight centre [**?**]) - GEOS Chem. GEOS-Chem is a global 3D model of atmospheric chemistry which is driven by the meteorology provided by NASA [GEOS-Chem, 2020]. Here the earth is split up into cubic sphere cells longitudally and latitudally, as well as vertically (Figure 1.3)[5]. Each one of these cells performs several purtubations of the chemistry within them, before any long-lived species are transported, and the process is repeated. If extracted separately a single one of these cells may be used to explore the sensetivity of different species for a range of input conditions. This is the bases of the atmospheric box model.

[5]This image is not from GEOS-Chem.

1.6.2 The box model.

In exploring the sensetivities of individual species within a simulation, it is possible to use a zero dimensional box model. This is in essence a single cell within the global structure, constrained in location and height (pressure). mechanism, integrator, etc.

It is then possible to take the many species, their rates or reaction and loss to produce a chemical mechanism detailing their properties in real life.

1.6.2.1 Chemical Mechanisms

Mechanisms are at the heart of every chemistry simulation. They are a mathematical representation of the possible reactions (and the rates at which these may occur) for every s

A note on model type, lifetime and mechanism size

Species such as OH and ... are very short lived and Ozone for example is long lived and can be transported

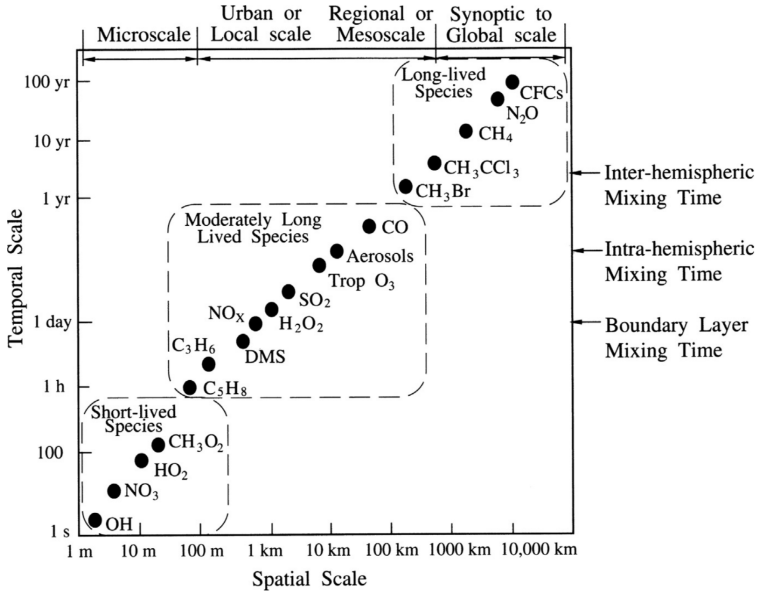


Figure 1.4: **Spatial and temporal scales of variability of atmospheric species.** Source: [Seinfeld and Pandis, 2016]

The atmosphere consists of thousands of species, with tens of thousands of reactions between them. These models represent real wodlc reactions In modelling these we can describe their rate of production and loss with respect to the species they react with.

1.6.3 Numerical integration

For example, it is possible to figure out how quickly each species in a reaction is changing if the reaction mechanism (the exact way it happe ns) and some simple data are known. This representation of how quickly the concentrations are changing is the same as a slope, or derivative. Integration allows us to find the actual change over time and not just how quickly the change is happening. Fo r example, given the following reaction, In a mechanism we are concerned with calculating how quickly a species changes within the chemical system. Taking the reaction of N_2O_5 (Equation 1.1) we can write the rate of change for each species over time (Equation 1.2)[6]. In integrating this equation, we are able to calculate the actual change in concentration (Equation 1.3) - this is the foundation of atmospheric models.

$$N_2O_5 \longrightarrow NO_2 + NO_3 \tag{1.1}$$

$$d[N_2O_5]/dt \longrightarrow d[NO_2]/dt + d[NO_3]/dt \tag{1.2}$$

$$\int d[N_2O_5]/dt \longrightarrow \int d[NO_2]/dt + \int d[NO_3]/dt \tag{1.3}$$

1.6.3.1 Non-Stiff Equations

Computational systems cannot integrate numbers analytically we rely on a series of computational algorithms. Since integration is the calculation of the area under a curve, the simplest of these

1.6.3.2 Numerically stiff equations (atmospheric chemistry)

Figure 1.4 shows the lifetimes of species can range between x orders of magnitude, similarly the componenets for each reaction (differential equation) evolve on significantly different timescales. This makes the atmospheric chemical mecahnism

1.6.4 The model development cycle

Scientific understanding is the product of many cycles of trial and error, Figure 1.5. In atmospheric chemistry we start with a hypothesis or a question, e.g. will changing X have a negative response on Y. We then construct a theoretical model to represent the chemsistry within. This chemistry is updated to reflect the rates and reactions that have been recorded in laboratory/chamber experiments. This cycle is then repeated until the model and real-world observations produce a comparable result.

[6]This is also known as the flux.

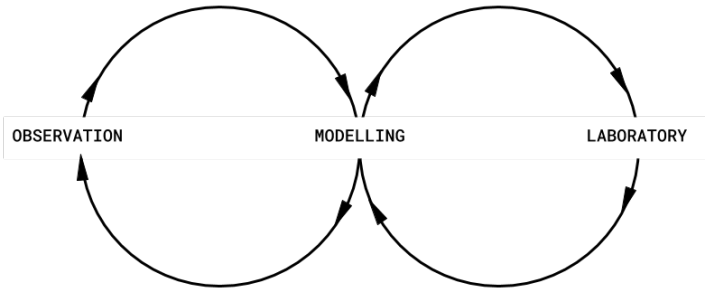


Figure 1.5: **The scientific development cycle.**This shows the iterative nature between modelling, observation and laboratory experimentation

ESM A series of box models.

1.6.5 The Dynamically Simple Model of Atmospheric Chemical Complexity

1.7 Thesis Layout

This thesis will explore a series of methods for describing and understaning the complex chemistry which may exist as part of an atmospheric chemistry mechanism. The mechanism used is a near-explicit representation of our foremost understanding of how gas phase chemistry in the troposphere reacts - the Master Chemical Mechanism, [**?**]. We begin by exploring the use of visualisation to convey complex scientific data (**??**). Next we apply this to the representation of species in a mechanism, and the relationships between them. To do this it is found that the node-link style graph format is the most beneficial, the use of which is then explored further (**??**). However in doing so, large complex network are shown to reach the limits of human cognition and visual representation. To overcome this a series of mathematic metrics are used to leverage our understanding of the species in a chemical network using graph theory (**??**). The use of computation to aid in graph analysis is further extended when graph clustering methods are applied as a method to group similar species within a chemical network (**??**). Finally in a bid towards the use of graph neural networks (see future work, **??**), a range of different chemical representations for machine learning are explored using a number of dimensionality reduction algorithms (**??**).

Bibliography

GEOS-Chem (2020). Geos-Chem Publications. http://acmg.seas.harvard.edu/geos/geos_pub.html.

Henderson-Sellers (2015). Climate Data Services | Nasa Center For Climate Simulation. https://www.nccs.nasa.gov/services/climate-data-services.

Höhne, N., den Elzen, M., Rogelj, J., Metz, B., Fransen, T., Kuramochi, T., Olhoff, A., Alcamo, J., Winkler, H., Fu, S., Schaeffer, M., Schaeffer, R., Peters, G. P., Maxwell, S., and Dubash, N. K. (2020). Emissions: World Has Four Times The Work Or One-Third Of The Time. *Nature*, 579(7797):25–28. http://dx.doi.org/10.1038/d41586-020-00571-x.

Hopkins, P. F. and Squire, J. (2018). The Resonant Drag Instability (Rdi): Acoustic Modes. *Monthly notices of the Royal Astronomical Society*, 480(2):2813–2838. https://academic.oup.com/mnras/article-pdf/480/2/2813/25498305/sty1982.pdf.

Organization, W. H. (2018). Who | Ambient Air Pollution: Health Impacts. https://www.who.int/airpollution/ambient/health-impacts/en/.

Peretó, J. (2011). *Oxygenic Photosynthesis*, pages 1209–1209. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-11274-4_1721.

Planavsky, N. J., Asael, D., Hofmann, A., Reinhard, C. T., Lalonde, S. V., Knudsen, A., Wang, X., Ossa Ossa, F., Pecoits, E., Smith, A. J. B., Beukes, N. J., Bekker, A., Johnson, T. M., Konhauser, K. O., Lyons, T. W., and Rouxel, O. J. (2014). Evidence For Oxygenic Photosynthesis Half A Billion Years Before The Great Oxidation Event. *Nature geoscience*, 7(4):283–286. https://doi.org/10.1038/ngeo2122.

Seinfeld, J. and Pandis, S. (2016). *Atmospheric Chemistry And Physics: From Air Pollution To Climate Change*. Wiley. https://books.google.co.uk/books?id=n_RmCgAAQBAJ.

Stevens, J. (2020). Airborne Nitrogen Dioxide Plummets Over China. https://earthobservatory.nasa.gov/images/146362/airborne-nitrogen-dioxide-plummets-over-china?fbclid=IwAR1z9jXZfY8xNZsCCRRo8Eor2hCjbNDIV70wXGOlzmNyFPkFBesURDCAwB4.

University of Zurich (2013). Great Oxidation Event: More Oxygen Through Multicellularity. *Science Daily*. https://www.sciencedaily.com/releases/2013/01/130117084856.htm.

Woo, M. (2018). Planet Formation? It'S A Drag. *Scientific American*. https://www.scientificamerican.com/article/planet-formation-its-a-drag/.

Appendices

Appendix A

Supplementary Mathematics

A.1 PCA

A.1.1 Statistics

Firstly we define the variance:

$$\sigma = \frac{\sum_{i=1}^{N}(X - \mu_X)(X - \mu_X)}{n - 1} \tag{A.1}$$

where X is the dataset, μ the mean and n the number of datapoints.

If we wish to then compare dataset X with dataset Y we may use the covarance:

$$cov(X, Y) = \frac{\sum_{i=1}^{N}(X - \mu_X)(Y - \mu_Y)}{n - 1} \tag{A.2}$$

For n distinct variables we may construct an $n \times n$ matrix containing $n!/(n-2)! \times 2$ different combinations of covariences:

$$C = \begin{pmatrix} \sigma_X & cov(X,Y) & cov(X,Z) & \cdots & cov(X,n) \\ cov(Y,X) & \sigma_Y & cov(Y,Z) & \cdots & cov(Y,n) \\ cov(Z,X) & cov(Z,Y) & \sigma_Z & \cdots & cov(Z,n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ cov(n,X) & cov(n,Y) & cov(n,Z) & \cdots & \sigma_n \end{pmatrix}$$

A.1.2 Matrices and Eigenvectors

An eigenvector is a vector $\mathbf{v}$, that when operated on by a given operator produces a scalar multiple of itself (Equation A.3) - this scalar multiple is called the eigenvalue λ. Eigenvectors can only be found for square matrices and are perpendicular to the matrix regardless of their dimension. A $n \times n$ matrix will produce n eigenvectors. Conventionally these are scaled to unity, which may be done by dividing the eigenvector by the pythoagorean distance of each element.

$$C\mathbf{v} = \lambda\mathbf{v} \tag{A.3}$$

An example of an eigenvector/value pair is shown in the following equations:

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = 4 \begin{pmatrix} 3 \\ 2 \end{pmatrix} \tag{A.4}$$

One property of the eigenvalue/eigenvector pair is that the square matrix acts as a transformation on the eigenvector. This means that we may treat the eigenvector as a direction from the origin, whose magnitude we can scale. The eigenvalue however remains scale independent and is the same value as before:

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 6 \\ 4 \end{pmatrix} = 4 \begin{pmatrix} 6 \\ 4 \end{pmatrix} \tag{A.5}$$

A.2 t-SNE

A.2.1 Student T distribution

Created by William Gosset and published under the pseudonym student [1] **?**.

The distribution consists of a family of continuous probability distributions which may be used when sample size is small and the standard deviaton is unknown. The curve itself resembles that of a normal distribution, just with a shorter amplitude and greater full width at half maximum (FWHM).

A.2.1.1 T-Score

Much like the z-score mentioned earlier [ref standardiz], t-scores also convert individual values to a standard form. This is genearlly ised when you dont know the population standard deviation (often due to having too few datapoints). At greater than 30 datapoints this resembles the equation of the z-score, and will often give you the same result.

$$t(x_i) = \frac{x_i - \mu_x}{S_{sample}/\sqrt{n}} \tag{A.6}$$

A.2.2 Kullback-Leiber (KL) divergence

KL divergence (also known as relative entropy) is a measure of distance between two distributions. It arises **?**.

`https://medium.com/syncedreview/kullback-leibler-divergence-explained-e358fbacf046`

[1] At the time Gosset was employed by Guinness Brewries in Dublin. This meant that chemists were forbidden from publishing their findings. After explaining that his mathematical and philosphical conculaions were of no use to competing brewries he was finally allowed to publish under the pseudonym 'student'. This was mainly to avoid difficulties with the rest of the staff.

Appendix B

Neural Network Activation Functions

B.1 Binary Step

.

This is a simple threshold function. If the input is above the threshold, the message is passed on. This makes it efficient, but unable to classify a single input into multiple categories. This can be likened to a yes|no decision tree.

$$f(x) = \begin{cases} 1, & \textbf{if } x < threshold \\ 0, & \text{otherwise} \end{cases} \tag{B.1}$$

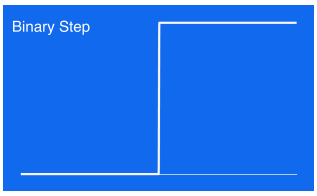


Figure B.1: Binary Step activation function.

B.2 Linear

.

This produces a signal proportional to the input multiplied by each neurons weight. It is an improvement over the step function as it allows for multiple outputs. It does however mean that we are unable to use backpropagation (gradient descent) to train the model. In adition to not being able to improve a model, all the layers in the neural network collapse into a single layer. This means that the final layer will always be a linear function of the first layer. This eliminates all the merits which may be gained from deep learning. A neural network with a linear activation function is simply a linear regression model.

$$f(x) = m(x) \tag{B.2}$$

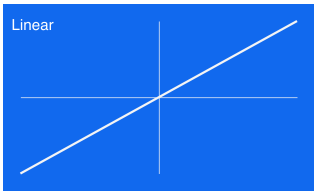


Figure B.2: Linear activation function.

B.3 Sigmoid / Logistic

.

The first of the non-linear activation functions. It has a smooth gradient providing smooth output values which are bound between 1 and 0, normalising the output of each neuron. The main disadvantage is that is falls foul the vanishing gradient problem - for extreme values of x there is close to no change in the prediction. This may result in either early termination of the training, or a slow training cycle in obtaining adequate precision. The activations is computationally expensive and the outputs are not zero centred.

$$f(x) = 1/1 + e^x \tag{B.3}$$

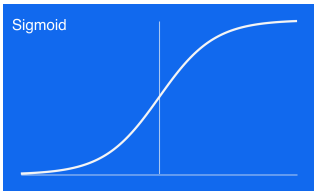


Figure B.3: Sigmoid activation function.

B.4 Hyperbolic Tangent

.

Much like the sigmoid function in both advantages and disadvantages. The hyperbolic tangent function provides a smooth curve which is zero centred. It is however computationally expensive and suffers from the vanishing gradient problem.

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{B.4}$$

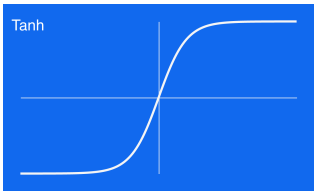


Figure B.4: Tanh activation function.

B.5 Rectified Linear Unit

.

A commonly used activation for large deep neural networks, due to its computational efficiency and quick convergence. It is non-linear although it appears like a linear function, and allows for back propagation. It does however suffer from the dying ReLU problem - when inputs tend to zero or below, the gradient of the function becomes zero and the network cannot perform backpropagation to learn.

$$f(x) = \begin{cases} 0, & \textbf{if } x < threshold \\ x, & \text{otherwise} \end{cases} \tag{B.5}$$

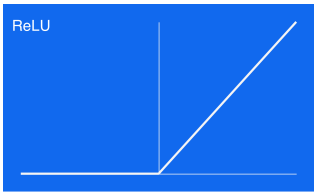


Figure B.5: ReLU activation function.

B.6 Swish

.

https://arxiv.org/abs/1710.05941v1 ***a new, self-gated activation function discovered by researchers at Google. According to their paper, it performs better than ReLU with a similar level of computational efficiency. In experiments on ImageNet with identical models running ReLU and Swish, the new function achieved top -1 classification accuracy 0.6-0.9% higher.***

$$f(x) = x/1 - e^{-x} \tag{B.6}$$

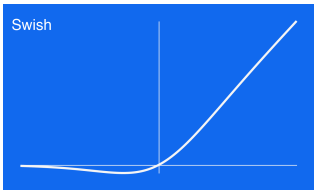


Figure B.6: Swish activation function.

B.7 A note on backpropagation

As it has not been explicitly explained before back-propagation is an algorithm used to train neural networks. The derivative (or gradient) of an activation function is important in the use of back propagation. Here the model weights are adjusted, and improved, by tracing back all the connections in network, suggesting an optimal weight of each neuron.

Appendix C

Miscellaneous

C.1 Correspondance with Mike Jenkin

Mike Jenkin 11th September 2019

Note on naming conventions in the CRI mechanism

The lumped or ‘‘common’’ species in the CRI mechanism are, by definition, used to represent a set of real species with different structures and properties. The criterion for lumping is the maximum number of NO-to-NO2 conversions (i.e. maximum number of ozone molecules) that the subsequent degradation can produce - and lumped species can therefore represent a large number of real species with different structures and properties.

In later expansions of the mechanism, the chemistry for species such as isoprene and terpenes defined intermediates that are representative of more restricted sets of real species. For these, it is possible to relate them to more restricted sets of MCM species that are the main contributors.

Although I tried to be logical in naming, the mechanism was developed over many years with little or no funding and may therefore not be fully transparent and foolproof throughout. However, I think quite a lot of the naming is logical, as expanded on below.

1) The numbers in most of the species names (the ‘‘CRI index’’) are the number of NO-to-NO2 conversions that can result from the subsequent OH-initiated NO-propagated chemistry. For radical termination products (e.g. hydroperoxides formed from RO2 + HO2 and nitrates formed from RO2 + NO), this is a grey area, and the number is therefore the same as that for the precursor RO2 radical. In these cases it is simply a convenient label.

2) There are a number of series of peroxy radicals, which are denoted RNxxO2, RIxxO2, RAxxO2, RExxO2, RUxxO2, RTNxxO2, RTXxxO2. These represent peroxy radicals with different structural features or formed from different types of precursor, as indicated below. Occasionally, extra peroxy radicals with the same CRI index are included by inserting a letter after the index (e.g. RNxxAO2) to increase flexibility of the mechanism. Peroxy radicals formed specifically from addition of NO3 to an alkene/diene are prefixed by ‘‘N’’.

RNxxO2: These were originally representative of peroxy radicals formed from linear or ‘‘n-‘‘ alkanes and their carbonyl products. They are also used for peroxy radicals formed from slightly-branched precursors (e.g. 2-methylhexane), and are formed as a convenient default intermediate with the correct CRI index in the latter stages of degradation of other precursor classes.

RIxxO2: These were originally representative of peroxy radicals formed from branched or ‘‘i-‘‘ alkanes and their carbonyl products, but tend to be used only for smaller branched precursors that can produce acetone as a major product from their subsequent degradation. This is because acetone is a particularly unreactive carbonyl, the formation of which can interrupt the ozone formation processes under typical regional-scale photochemical episode conditions in north-west Europe.

RAxxO2: These peroxy radicals are formed from the addition of OH to aromatic compounds, and are complex bicyclic structures containing a peroxide bridge (e.g. like BZBIPERO2 in the MCM).

RExxO2: These peroxy radicals are formed from ether degradation, and allow the formation of unreactive formate ester products to be represented.

RUxxO2: These peroxy radicals are formed from degradation of conjugated dienes (currently only isoprene and 1,3-butadiene). Those formed initially (e.g. RU14O2) contain allyl functionalities (i.e. a specific unsaturated linkage), although the terminology is also used for some peroxy radicals formed from subsequently-formed unsaturated products.

Related to this, the species CRU14O2 and TRU14O2 in the EMEP variant of CRI v2.2 (described in https://doi.org/10.1016/j.atmosenv.2019.05.055) were specifically introduced to represent the cis- and trans- isomers required for the Peeters (LIM) reaction framework. CRU14O2 represents CISOPAO2 and CISOPCO2 in MCM v3.3.1 and TRU14O2 represents ISOPAO2 and ISOPCO2 in MCM v3.3.1. However, CRI v2.2 itself uses a different approach where the chemistry is represented by a conditions-dependent rate coefficient for the single peroxy radical, RU14O2.

RTNxxO2: This terminology is used for peroxy radicals formed from monoterpenes containing an endocyclic double bond. This is currently limited to α-pinene in CRI, although the original idea was that the mechanism could be used as a surrogate for other endocyclic monoterpenes by simply adding new sets of initiation reactions.

RTXxxO2: This terminology is used for peroxy radicals formed from monoterpenes containing an exocyclic double bond. This is currently limited to β-pinene in CRI, although the original idea was that the mechanism could be used as a surrogate for other exocyclic monoterpenes by simply adding new sets of initiation reactions.

Finally, the species DHPR12O2 in CRI v2.2 is a peroxy radical containing two hydroperoxy groups. Again, it is required for representation of the Peeters (LIM) mechanism, and is representative of the species C536O2 and C537O2 in MCM v3.3.1 (these species being referred to as ‘‘di-HPCARPs’’ by Peeters et al., 2014: https://doi.org/10.1021/jp5033146).

3) Hydroperoxides formed the reactions of HO2 with the above peroxy radicals have ‘‘OOH’’ in place of ‘‘O2’’. Nitrates formed the reactions of NO with the above peroxy radicals have ‘‘NO3’’ in place of ‘‘O2’’.

4) There are a number of series of carbonyl compounds, which are denoted CARBxx, UCARBxx, UDCARBxx, TNCARBxx and TXCARBxx.

CARBxx: These are used to represent carbonyls and hydroxycarbonyls. Occasionally, extra carbonyls/hydroxycarbonyls with the same CRI index are included by inserting a letter after the index (e.g. CARBxxA) to increase the flexibility of the mechanism.

Related to this, the species DHPCARB9 in CRI v2.2 is a carbonyl containing two hydroperoxy groups. Again, it is required for representation of the Peeters (LIM) mechanism, and is representative of the species DHPMEK and DHPMPAL in MCM v3.3.1 in MCM v3.3.1.

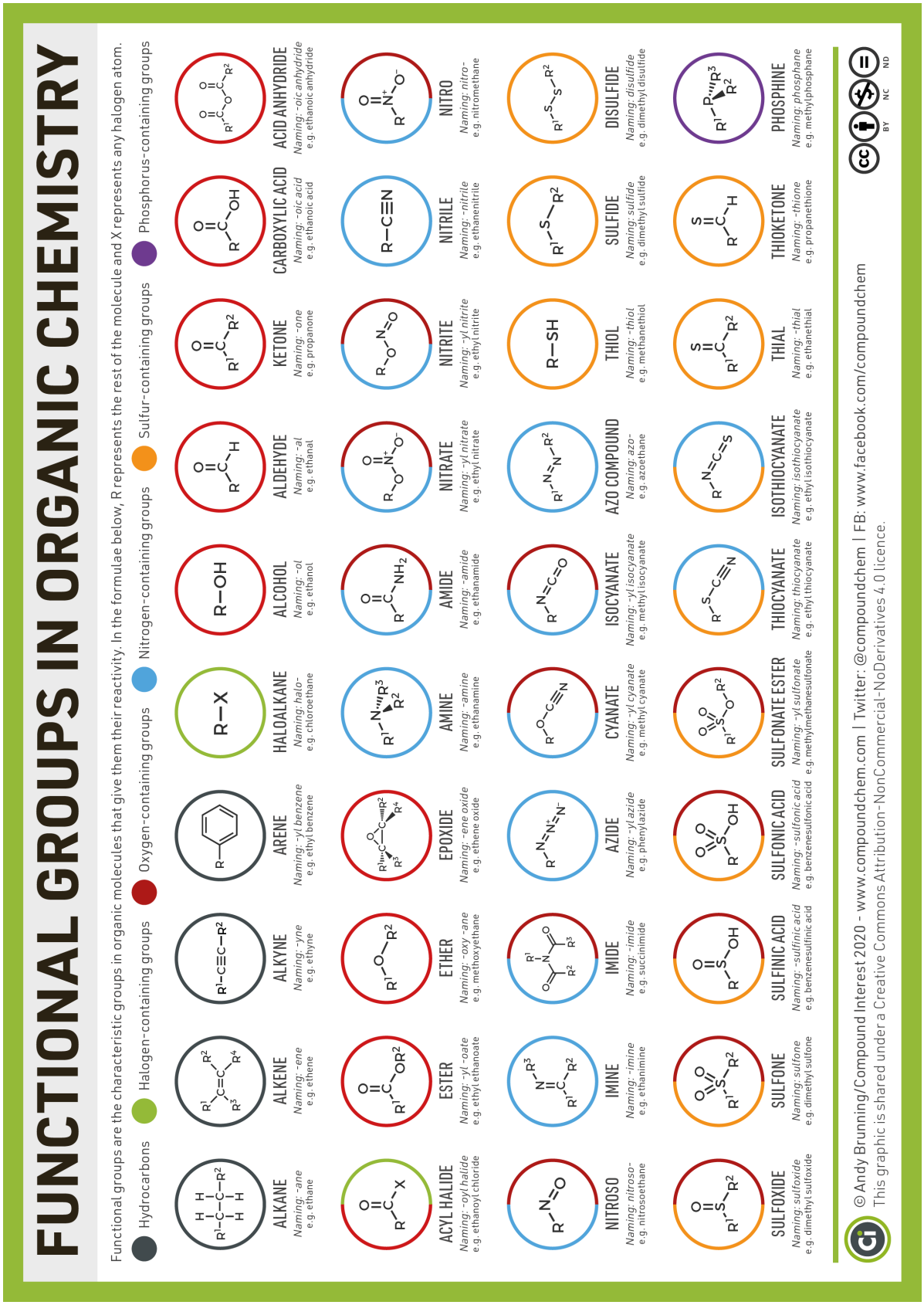
UCARBxx: This terminology is used for unsaturated carbonyls/hydroxycarbonyls, formed for example from isoprene (although one of the main ones, UCARB10, has been ‘‘unlumped’’ into MVK and MACR in the EMEP CRI v2.2 variant).

Related to this, the species HPUCARB12 in CRI v2.2 is an unsaturated carbonyl containing a hydroperoxy group. Again, it is required for representation of the Peeters (LIM) mechanism, and is representative of the species C5HPALD1 and C5HPALD2 in MCM v3.3.1.

UDCARBxx: This terminology is used for unsaturated dicarbonyls, formed from aromatics.

TNCARBxx and TXCARBxx: This terminology is used for carbonyl compounds, formed from monoterpenes with endocyclic and exocyclic double bonds, respectively.

C.2 Functional Groups



Source:?

Appendix D

Chapter Keywords

This section uses the Term Frequency Inverse Document Frequency to determine the keywords of each chapter - a techneque which has been described in **??** and **?**. Text size corresponds to the importance of each word.

D.1 Introduction

AIR SPECIES POLLUTION ATMOSPHERE EARTH ATMOSPHERIC MODEL CHEMISTRY EQUATION DT CHEMICAL OZONE REACTION MECHANISM CLIMATE OXYGEN NUMERICAL EQN DEVELOPMENT HUMAN CYCLE SYSTEM NITROGEN GAS AGO CONCENTRATIONS GCM TIMESCALES ALSO POSSIBLE EVENTUALLY HUMANITY MODELLING DEVCYCLE INDOOR AMBIENT PLANET INT POLLUTANTS HEALTH CHINANOX LIFE EFFECTS ESM RATES MANY CHANGE SOURCE MODELS REPRESENTATION INCREASE RANGE CHANGING COMPLEX UNDERSTANDING GRAPH EXAMPLE YEARS SCIENTIFIC DATA QUALITY BOX GLOBAL OH REACTIONS CELLS SERIES QUICKLY BECOME INTEGRATION TIME PRODUCTION GEOSCHEM NOX THESIS ACT TRAGEDY RADIATION FLAW LIVED DISEASE DRAG HOMO RESPIOTORY WEATHER DEATHS BILLION INDUSTRIAL SE PREMATURE ESMS EXPERIMENTS LIFETIME CHINA ACTUAL LIMITS TRANSFER FLUSHLEFT WHODATA OZONEROLE

D.2 Applying Visual Analytics to the Atmospheric Chemistry Network

GRAPH NODES LAYOUT NODE GRAPHS EDGES SPECIES EDGE LAYOUTS REPRESENTATION USING DISTRIBUTION ALGORITHM CONFLUENT DENSITY DESIGN FORCEDIRECTED NETWORK MECHANISM MERCATOR ANGLE REACTIONS INFORMATION MAY CHEMISTRY MERC STRUCTURE VISUALISATION POSSIBLE FORCE CHEMICAL DATA LINKS MCM APHH SEMANTIC CM DIFFERENT DEGREE ALTHOUGH TSNET OPENORD SHOWS USER ONE REPRESENT BUTANE DRAWING HU ROUTING CROSSING BEZIER ITEMIZE REPRESENTING NUMBER SYSTEM EXAMPLE VISUAL BUNDLING YIFAN ATLAS SHAPE ADDITION AREA PROCESS MANY POINTS BEIJING BEST FORCEATLAS COLOR QUADTREE ORTHOGONAL DIRECTED HIGH PRIMARY SET DESCRIBED ADDITIONAL SELECTION ITEMS REDUCE HOWEVER METHODS NETWORKS SINCE SYNTACTIC GENERAL CURVES GENERATED RESOLUTION STUDY FOUND LARGE ALSO ENERGY DIRECTION CARBON USEFUL SIZE

D.3 Computational Learning, Visualisation and Clustering:

SPECIES PCA DATA CLUSTERS DATASET TSNE SMILES STRUCTURE ALGORITHM GRAPH USING VEC METHODS RANDOM NODE POINTS GROUPS ALGORITHMS DR NUMBER REDUCTION CLUSTERING VERB CLUSTER DIMENSIONALITY AUTOENCODER DIFFERENT SILHOUETTE EQUATION FUNCTIONAL LINEAR DIMENSIONS SET MATRIX VECTOR INPUT ORIGINAL FEATURES COLOURS DISTRIBUTION TABLES SUBIMPORT QUANTUM PRINCIPAL PROBABILITY ACTIVATION TEX MAY POSSIBLE REPRESENTING SINCE MCM METHOD COMPONENT ST FINGERPRINTS RESULT TWO INPUTS FEATURE DESCRIBED STRING MOLECULAR ALTHOUGH CONSTRUCTION OUTPUT ONE CHEMICAL NEW DECISION KEYS PARAMETERS BEST COEFFICIENT NETWORK SPACE RANGE EXAMPLE COLOUR CHEMISTRY HOWEVER REDUCED NODES VALUE NONLINEAR RESULTS GRADIENT GAUSSIAN GREEDY SYSTEM EQN FOUR WORKS STEP MAPPING COMPARING DISTANCE SEVERAL PROPERTIES LEARNING