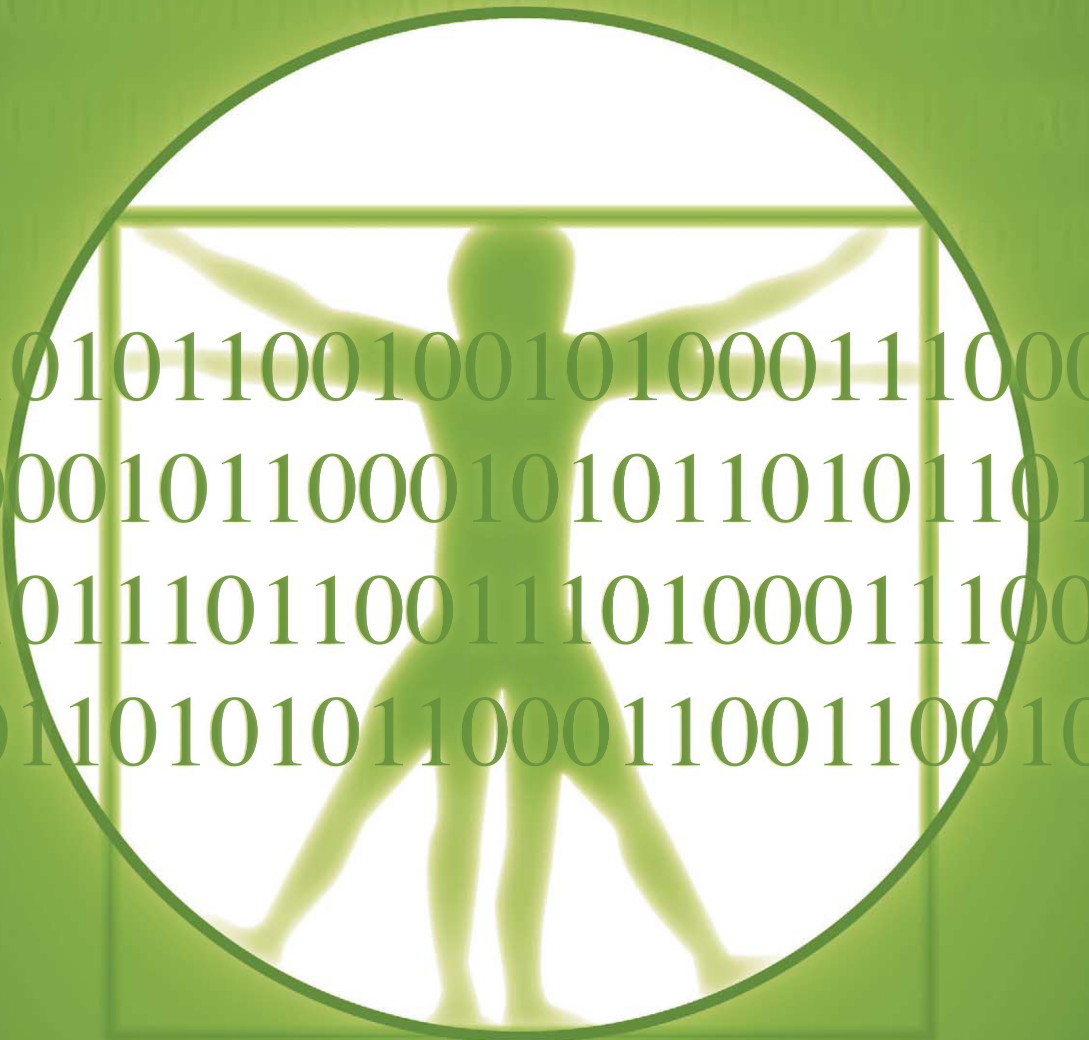
Ambient Intelligence and Smart Environments **24**

Agriculture and Environment Perspectives in Intelligent Systems



Edited by

Andrés Muñoz

Jaehwa Park

IOS Press

Agriculture and Environment Perspectives in Intelligent Systems

The eventual aim when applying digital technologies in agriculture is to replace or reduce the human labor required for agricultural production. Large amounts of heterogeneous data are essential for integration studies of automated agriculture, and the digitalization of agriculture is helping to fulfill the demand for this data, but management of the data gathered presents its own challenges. That is where the Intelligent Environment (IE) paradigm comes into play to guide the design of the systems, techniques and algorithms able to analyze the data and provide recommendations for farmers, managers and other stakeholders.

This book, Agriculture and Environment Perspectives in Intelligent Systems, is divided into 5 chapters. Chapter 1 explores the use of intelligent systems in Controlled Environment Agriculture (CEA) facilities; Chapter 2 reviews the adoption of intelligent systems in the research field of biomonitoring; Chapter 3 proposes an intelligent system to acquire and pre-process data for precision agriculture applications; Chapter 4 illustrates the use of intelligent algorithms to make more efficient use of scarce resources such as water; and Chapter 5 focuses on the generation of intelligent models to predict frosts in crops in south-eastern Spain.

There is still a need to bridge the gap between the needs of farmers, environmental managers and stakeholders and the solutions offered by information and communication technology. This book will be of interest to all those working in the field.

ISSN 1875-4163 (print)
ISSN 1875-4171 (online)
ISBN 978-1-61499-968-3 (print)
ISBN 978-1-61499-969-0 (online)

9 781614 999683

AGRICULTURE AND ENVIRONMENT PERSPECTIVES IN INTELLIGENT SYSTEMS

Ambient Intelligence and Smart Environments

The Ambient Intelligence and Smart Environments (AISE) book series presents the latest research results in the theory and practice, analysis and design, implementation, application and experience of *Ambient Intelligence* (AmI) and *Smart Environments* (SmE).

Volume 24

Recently published in this series

Vol. 23. I. Chatzigiannakis, Y. Tobe, P. Novais and O. Amft (Eds.), Intelligent Environments 2018 – Workshop Proceedings of the 14th International Conference on Intelligent Environments
Vol. 22. C. Analide and P. Kim (Eds.), Intelligent Environments 2017 – Workshop Proceedings of the 13th International Conference on Intelligent Environments
Vol. 21. P. Novais and S. Konomi (Eds.), Intelligent Environments 2016 – Workshop Proceedings of the 12th International Conference on Intelligent Environments
Vol. 20. W. Chen et al. (Eds.), Recent Advances in Ambient Assisted Living – Bridging Assistive Technologies, e-Health and Personalized Health Care
Vol. 19. D. Preuveneers (Ed.), Workshop Proceedings of the 11th International Conference on Intelligent Environments
Vol. 18. J.C. Augusto and T. Zhang (Eds.), Workshop Proceedings of the 10th International Conference on Intelligent Environments
Vol. 17. J.A. Botía and D. Charitos (Eds.), Workshop Proceedings of the 9th International Conference on Intelligent Environments
Vol. 16. M. d’Angelantonio and J. Oates (Eds.), Is Ambient Assisted Living the Panacea for Ageing Population?
Vol. 15. J. Mintz et al. (Eds.), Touching the Future Technology for Autism? – Lessons from the HANDS Project
Vol. 14. M. Bhatt and H.W. Guesgen (Eds.), Situational Awareness for Assistive Technologies

ISSN 1875-4163 (print)
ISSN 1875-4171 (online)

Agriculture and Environment Perspectives in Intelligent Systems

Edited by

Andrés Muñoz

Polytechnic School, Catholic University of Murcia (UCAM), Murcia, Spain

and

Jaehwa Park

Chung-Ang University, Seoul, South Korea

IOS
Press

Amsterdam • Berlin • Washington, DC

ISBN 978-1-61499-968-3 (print)
ISBN 978-1-61499-969-0 (online)
doi: 10.3233/AISE24

Publisher
IOS Press BV
Nieuwe Hemweg 6B
1013 BG Amsterdam
Netherlands
fax: +31 20 687 0019
e-mail: order@iospress.nl

For book sales in the USA and Canada:
IOS Press, Inc.
6751 Tepper Drive
Clifton, VA 20124
USA
Tel.: +1 703 830 6300
Fax: +1 703 830 2300
sales@iospress.com

PRINTED IN THE NETHERLANDS

Preface

For the last decade, intensive and integrative studies have been performed by applying information technology to the research fields of agriculture and environmental sciences. This research trend aims to improve the agricultural process using information technology and at the same time making this process more sustainable. Applying sensor technology and combining information technologies creates the opportunity to obtain more precise information on agriculture processes and environmental phenomena. Indeed, the Internet of Things paradigm opens new frames to collect large amounts of agricultural and environmental data through different kinds of sensor networks [1].

The agricultural control starts with analyzing the current situation using chronological data. Traditionally, growers observe plants and detect problems based on their experience and symptoms of diseases. Most of such cases, the detection critical point is usually late in sense of optimization. An automatic method to detect symptoms and provide an early warning would at least help to improve the efficiency of agricultural processes. The correlations between environmental factors are the keys for evaluating agricultural characteristics. In this sense, a large amount of heterogeneous data is essential for integration studies of automated agriculture. There is a high demand for a large quantity of accurate spatiotemporal data for automated agriculture. Whilst the digitalization of the agriculture is helping to fulfill this demand, new challenges are posed related to the intelligent management of the data gathered from sensors to help farmers and environmental managers make decisions. It is in these challenges where the Intelligent Environment (IE) paradigm comes into play to guide the design of systems, techniques and algorithms able to analyze the data and provide recommendations that can be put into practice by farmers, managers and other stakeholders.

Some examples of applications of IE in this research field range from smart farming (see for example [2–4]) to intelligent applications for ecological disaster management (see for example [5] where a social sensor is developed for detecting floods through natural language processing (NLP)). Specific examples of IE applications in agriculture can be found for resource allocation and decision making in agricultural systems [6,7], farming system design [8] or data fusion for characterizing key agricultural system attributes [9]. Likewise, other IE proposals for environmental management can be found in different areas such as waste management [10] or environmental assessment and rehabilitation [11]. It is also noteworthy the increasing number of conferences and workshops devoted to the development of IEs in agriculture: approximately 30 events were found in the WikiCFP webpage in 2018 tagged with “agriculture” and “intelligent environments” as keywords, as for example the International Workshop on Intelligent Systems for Agriculture Production and Environment Protection (ISAPEP).

From our experience, we believe there is still a need for bridging the gap among the needs of farmers, environmental managers and stakeholders in the agriculture and environmental sectors and the solutions offered by the Information and Communication Technology field. The eventual goal of applying IE to agriculture is to replace, or significantly reduce, the human labor for the agricultural production. Most of contemporary research on Artificial Intelligence (AI) technology focuses on finding methods to simulate human perception, cognition and reasoning process. However, still a long way to reach the human ability remains. When the AI technology reaches maturity, the dramatic transformation on devices, equipment and machinery of agriculture production will begin. The motivation of editing this book was to foster the collaboration among researchers from computer science, agriculture engineering and environmental science fields who are in a privileged position to provide new and valuable insights on intelligent applications for agriculture and environmental management systems. As a result, this book presents several reviews and examples of the latest developments in the application of IEs to agriculture and environmental sciences through 5 chapters.

The editors of this book would like to thank the authors who contributed articles to this initiative. The diversity of the topics included in the chapters showcases the relevance in the research and development of Intelligent Environments applied to agriculture and environmental problems. We look forward to a future where IE will make a difference in creating efficient and really helpful systems for farmers, environmental managers and the entire society.

Outline of the book

Chapter 1 explores the use of intelligent systems in Controlled Environment Agriculture (CEA) facilities as an alternative to offer low cost sensing systems to gather data on how plants evolve in these facilities. At the moment, data collected in CEAs are restricted to recording the general environmental control responses. In this chapter the author relates his experiences within a farm company when developing more fine-grained methods to gather data, especially focused on those related to growing methods such as hydroponics or aeroponics and on intelligent techniques for measuring nutrients.

Chapter 2 reviews the adoption of intelligent systems in the research field of biomonitoring. In this chapter is described the evolution of environmental information systems and their application in biomonitoring, starting from the use of biosensors to the analysis of current environmental information systems to the introduction of intelligent techniques and methods in these systems. Several examples of applications of intelligent biomonitoring systems are analyzed, from systems to monitor algae to applications for monitoring fish. The authors also include a discussion on the limitations and future perspectives of intelligent biomonitoring systems.

Chapter 3 proposes an intelligent system to acquire and preprocess data for precision agriculture applications. It is composed of two modules, the first aimed to measure different environmental factors on the target plants and the second for refining time-series raw data previously collected in order to reduce noise in the data. The main goal of this system is to facilitate the acquisition of different type of data to farmers, since it allows any analog sensor to be connected with minimal effort of calibration. In this manner, the system is presented as a user-friendly device for multi-data collection which may help in the creation of big data sources for new applications.

Chapter 4 illustrates the use of intelligent algorithms to make a more efficient use of a scarce resource such as water. These algorithms predict the intake of pharmaceutical contaminant elements such as diclofenac in reclaimed water for irrigation of lettuces. As a result, this proposal allows farm managers to automatically control the quality of the reclaimed water and effectively use it when it is safe. The authors compare several machine learning algorithms to this end, proving that the Random Forest algorithm can reach a 97% of fitness when predicting the presence of contaminants in the water.

Chapter 5 focuses on the generation of intelligent models to predict frosts in crops in Southeastern Spain. These models not only use data from physical sensors placed in the land or trees, but they also rely on the utilization of open data from weather stations.

By using different rule and decision trees techniques, the authors are able to identify the most relevant features for predicting frosts, such as dew point, vapor pressure deficit and maximum relative humidity. The results show that the proposed model is able to predict a frost with 98% of confidence and obtain an error of less than 0.5ºC. in the prediction of the minimum temperature.

Andrés Muñoz[1] and Jaehwa Park

References

[1] Ray PP. Internet of things for smart agriculture: Technologies, practices and future direction. *Journal of Ambient Intelligence and Smart Environments* **9**(4). 2017, 395–420. doi: 10.3233/AIS-170440

[2] Aziz IA, Ismail MJ, Haron NS, Mehat M. Remote monitoring using sensor in greenhouse agriculture. In *2008 International Symposium on Information Technology*, Vol. 4, IEEE. 2008, 1–8. doi: 10.1109/ITSIM.2008.4631923

[3] Aquino-Santos R, González-Potes A, Edwards-Block A, Virgen-Ortiz RA. Developing a new wireless sensor network platform and its application in precision agriculture. *Sensors* **11**(1). 2011, 1192–1211. doi: 10.3390/s110101192

[4] Muñoz A, Soriano-Disla JM, Janik LJ. An Ontology-Based Approach for an Efficient Selection and Classification of Soils. In *Analide C. and Kim P. (editors), Intelligent Environments 2017: Workshop proceedings of the 13th International Conference on Intelligent Environments*, Vol. 22, IOS Press. 2017, 69–79. doi: 10.1016/j.biosystemseng.2018.11.016

[5] Periñán-Pascual C, Arcas-Túnez F. Detecting environmentally-related problems on Twitter. *Biosystems Engineering* **177**. 2019, 31–48. doi: 10.1016/j.biosystemseng.2018.10.001

[6] Jiang Y, Hao K, Cai X, Ding Y. An improved reinforcement-immune algorithm for agricultural resource allocation optimization. *Journal of computational science* **27**. 2018, 320–328. doi: 10.1016/j.jocs.2018.06.011

[7] Kaim A, Cord AF, Volk M. A review of multi-criteria optimization techniques for agricultural land use allocation. *Environmental Modelling & Software* **105**. 2018, 79–93. doi: 10.1016/j.envsoft.2018.03.031

[8] Robert M, Dury J, Thomas A, Therond O, Sekhar M, Badiger S, Ruiz L, Bergez JE. CMFDM: A methodology to guide the design of a conceptual model of farmers' decision-making processes. *Agricultural Systems* **148**. 2016, 86–94. doi: 10.1016/j.agsy.2016.07.010

[9] Jiayu Z, Shiwei X, Zhemin L, Wei C, Dongjie W. Application of intelligence information fusion technology in agriculture monitoring and early-warning research. In *2015 International Conference on Control, Automation and Robotics,* IEEE. 2015, 114–117. doi: 10.1109/ICCAR.2015.7166013

[10] Anagnostopoulos T, Zaslavsky A, Kolomvatsos K, Medvedev A, Amirian P, Morley J, Hadjieftymiades S. Challenges and opportunities of waste management in IoT-enabled smart cities: a survey. *IEEE Transactions on Sustainable Computing* **2**(3). 2017, 275–289. doi: 10.1109/TSUSC.2017.2691049

[11] Sun AY, Zhong Z, Jeong H, Yang Q. Building complex event processing capability for intelligent environmental monitoring. *Environmental Modelling & Software* **116**. 2019, 1–6. doi: 10.1016/j.envsoft.2019.02.015

[1] Andrés Muñoz would like to thank the Spanish Ministry of Economy and Competitiveness for its support under the project TIN2016-78799-P (AEI/FEDER, UE).

This page intentionally left blank

Contents

This page intentionally left blank

Agriculture and Environment Perspectives in Intelligent Systems
A. Muñoz and J. Park (Eds.)
IOS Press, 2019

doi:10.3233/AISE190002

Controlled Environment Agriculture: Using Intelligent Systems on the Next Level

Jonathan LODGE
CEO, City Farm Systems Ltd and Robotic Farm Systems Ltd, UK

Abstract. The need to increase the supply of food for larger and increasingly urbanised populations has created what many see as an entirely new sector based around reducing the distance salad crops travel. This usually involves growing delicate crops in enclosed spaces where the environment is controlled to give what is thought of as ideal growing conditions. In reality, since plants have evolved over many millennia, the process of growing a plant needs many of the variations nature delivers. We are only just beginning to learn what this means for the human diet. It is now apparent growing conditions cannot be changed substantially without changing some known and unknown qualities. This chapter looks first at what is being offered by those investing in Controlled Environment Agriculture (CEA) and then how, if done properly, the use of Intelligent Systems can unlock much greater benefits and then go further by informing some aspects of traditional agriculture. CEA needs to look further at why plants evolved as they did and use Intelligent Systems to build that knowledge and make effective use of it. Some aspects of traditional agriculture are using technology to deliver far greater granularity of data than those who grow in environments more usually found in a laboratory. The overwhelming majority of CEA practioners talk of data but then do little more than record environmental setpoints and rely on measuring Electrical Conductivity (EC) as a proxy for assessing nutrient levels. This chapter shows how the current array of low cost sensing options offer enormous improvement for growers and researchers. Introducing truly intelligent use of contemporary technology can create valuable outcomes to improve many areas of agriculture.

Keywords. Urban Farming, Vertical Farming, Urban Agriculture, Controlled Environment Agriculture (CEA), Hydroponics, Aeroponics, Aquaponics.

1. Introduction

As the world's population increases the proportion living in urbanised areas is increasing rapidly. The first industrial revolution both enabled and relied on an agricultural revolution. Ever since then developed countries have seen populations move from the land and into cities. Many cities are now struggling to cope with the levels of traffic. Increasingly, as affluent cities use of cars increases along with congestion and related pollution, the average speed of traffic is returning to that of horse drawn carriages and bicycles. Many city planners now make efforts to be more sustainable and planet friendly by reducing traffic and cleaning up the environment. New areas are often planned without many roads. However, the need to feed these communities remains and supply chains are ever more complex.

The need to improve food supply chains for cities is an issue found around the world and some are attempting to reduce the distance delicate crops travel. This distance is often referred to as 'food miles'. In his book *The Vertical Farm*: *Feeding the World in* *the 21st Century,* Dickson Despommier [1] suggested a future for horticulture where food would be grown in multiple layers in city buildings. Many academics and some entrepreneurs have followed his thinking. Unfortunately, this 'food at any cost' approach does not look at commercial realities. More often than not those who follow this path use 'hydroponics' to grow a restricted range of crops using the same business model as most traditional farmers; growing at maximum capacity and then hoping to sell all they can harvest. This is a major drawback of current farming systems that concentrate on field based yield and do little to recognise the proportion of a harvest that reaches a consumer's plate. Far more needs doing to measure delivered quantities. For short cycle crops (and livestock farming) we need to look much more at demand side economics unlike arable farmers who only need to look at supply side economics.

Having paid to exclude natural sunlight they have to pay even more to create light artificially that rarely delivers the complete and varying spectrum a plant needs. The trend for using blue and red LED lighting has been claimed by some to be a solution seeking a problem to solve. They argue that, as leaves are green, that is the part of the spectrum they are reflecting and, therefore do not need. It can be shown that this is what restricts many to grow leafy greens as they do not offer parts of the sun's radiation spectrum that trigger many growth characteristics. The human eye has specialist green light receptors whereas other species may see beyond and or only part of the humanly visible spectrum. It must be recognised this issue is far more complex than many realise.

This chapter seeks to show how making better use of Intelligent Systems a Controlled Environment Agriculture (CEA) facility can move on from loss making examples and offer an economically viable alternative. Many claim to use intelligent systems but in reality their approach is restrictive and locks them into the traditional business model. It can be argued that the data they collect is little more than recording the environmental control responses of a whole, homogenised environment. Programming temperature changes through a day takes this a stage further and may create what is believed to be ideal growing conditions but does not encourage some of the qualities found in the same crop grown in more natural conditions.

Invariably the nutrient rich water used in Hydroponics is monitored by measuring Electrical Conductivity (EC) as a proxy for the amount of nutrients present [2]. Recent developments mean the cost of measuring individual nutrients is coming down to a level where it should be replacing EC. This is the starting point where it can be seen CEA should not be thought of as a completely separate specialism and can both learn from and inform many areas of traditional agriculture.

Increasingly there is evidence to show there are nutrients that may not be essential for healthy plant growth but are what creates the taste and flavour qualities we like together with the ability to deliver trace elements necessary for the diet of a healthy human.

In essence CEA is the logical extension of growing plants in a protected environment such as a greenhouse where growing seasons could be extended with warmer conditions. Greenhouses are often referred to as 'glasshouses' as all early enclosed growing facilities used glass to allow as much natural sunlight in as possible. All early adopters found they needed to ventilate glasshouses to reduce overheating in summer and to bring fresh air in. It was soon realised that, without bringing fresh air, carbon dioxide levels dropped which caused an unexpected reduction in growth rates. Experience showed several environmental aspects needed adjustment to aid growth or prevent unwanted mildew and disease. The majority of CEA practioners now rely on simply extending the control of these environmental parameters with many believing that artificial light can be similarly controlled. Unfortunately too many practioners have come from engineering backgrounds and have not engaged sufficiently with crop researchers. Who would have thought a plant reacts to a circadian rhythm? See TiMet an Horizon 2020 funded research project [3].

Extensive research by the author's business shows there are many ways intelligent systems can go beyond the (relatively) simple management of a greenhouse and automate the whole growing cycle. Making the most of this goes further and offers a new business model that can help feed a modern city. Part of the latest agricultural revolution (often referred to as Agriculture 4.0 and relating to Industry 4.0) needs to adopt some of the principles of current industrial thinking such as that started by William B Smith Jr. [4] to make a city smart. William B Smith Jr came up with the core thinking that evolved into Lean Six Sigma as the avoidance of unnecessary cost when he was at Motorola in 1980. This is largely around minimising cost and effort required to achieve desired output qualities. Latest developments show how using this approach in some CEA systems can offer significantly improvements and enhance research programs for both crops and growing media.

Section 2 discusses the background of the CEA sector. Section 3 reviews the basic types of growing methods used in CEA. Section 4 discusses nutrient needs in these environments. Finally, Section 5 is all about how applying Intelligent Systems can make CEA more effective and enable both to offer benefits for some areas of traditional agriculture.

2. Background

2.1. Vertical Farming

This is a term that many use but one the author dislikes intensely for a simple reason. It was first used by Dickson Despommier [1]. It can be argued however that he missed the importance of city economics. Obviously urban farmers talk of growing more than traditional agriculture. Despommier suggested in his book that growing 100 times as much per unit of area would be achieved and that, with more research, much greater yield levels should be possible. Some growers now talk of a figure near 150 with more being targeted. At first glance those figures are amazing but they pale into insignificance when looking at the costs practitioners experience when growing in a city. The cities that most need help with their food supply chains are the busiest. During the last decade there have been some industrial buildings in quiet American cities available with rent charges that are unsustainably low. In some instances they can be so low (even for new buildings) that it is hard to see how the construction of the building was financed. In a busy city, rents can be far too high to devote the space to growing food when they can attract high rents as living or working space. In the UK rent costs for the cheapest commercial space (such as railway arches) in or near London are 20,000 times as much as for agricultural land and need substantial remedial work to make them suitable for growing food. Many building rents in London are 40-50,000 times that for agricultural land. So looking again at the finances it can be seen that most urban farmers are paying several hundred times as much for their growing capacity as current competitors – and that's before they install their equipment.

The author's lightbulb moment was triggered when needing to buy groceries on the way home from work. He found himself stuck in traffic for several minutes behind the retailer's large truck only yards from the store. Once inside he found the fresh produce shelves were bare. Not only was one of the most expensive vehicles on the road achieving nothing but the store was losing sales while the produce was getting nearer to the end of its shelf life. Thinking about the situation there were many points of failure that extended far beyond the obvious.

Early on the author had a meeting with Professor Marian Rizov [5] who suggested the need to reduce the 'minimum efficiency scale' (MES). Looking at rural commercial greenhouse producers it can be seen they rely on scale. To make a profit they rely on large scale harvesting and delivery by the truck load. This locks them into the low margin monoculture supply chains which depend on distribution centres for onward delivery to individual outlets that only need a few boxes of lettuce once or twice a week. In 2013 the UK retailer Tesco was widely reported as suggesting two thirds of bagged salads ended up as waste [6&7]. It was reported by some that part of this was on farm, some in transit or instore and the rest was in the home. Food waste is a major priority for UK based charity Waste and Resources Action Programme (WRAP) who have created articles and initiatives such as those titled *The Food Waste Reduction Roadmap* and *The Business Case for Reducing Food Waste in Restaurants* [8].

Considering that CEA generally concentrates on produce that costs more to deliver than to grow more attention should made that much of what has been grown will never be consumed. Knowing this and it can be seen that the costs of distribution were pointless. This is a simple statistical approach. Let's consider the reality behind those statistics. It is safe to assume this was expressed as a percentage of harvestable produce. Most people in business and all statisticians know the standard curve of normal distribution or bell curve. If we apply this to the point at which a growing crop should be harvested, we can see that a major grower would need to grow more than enough to fulfil a large order. This means that if the grower is only growing for large orders then the produce ready for harvest early and late is unwanted. By the time one counts the crops that do not make a retailer's acceptable qualities for harvest (often a random size requirement) the amount of overall crop wasted is enormous. Now consider what happens for specific events. Over a holiday weekend a retailer needs to allow for extremes. For such a weekend they have to make longer term buying decisions than normal and so plan for the most likely selling pattern. If the holiday weekend weather takes a sudden turn for the worse salad stock levels will be higher than needed and the amount wasted could be much greater. Conversely if the weather is better than expected the stores will have empty shelves and lose sales. The percentage of waste is therefore understated in most scenarios. The headline figure is often far higher or only expressed as a fraction of what could have been sold. Without mentioning the obvious difficulties, the UK voted to increase there have long been regular difficulties at the Channel Tunnel. A large amount of food crosses the channel and there are regular failures in either the infrastructure or the manning of them. When shelf life is measured in hours rather than days such delays can consume much of the shelf life and it is not unknown for whole truckloads to be rejected causing 100% wastage. Having a more flexible supply chain can buffer harvests to match short term demand changes and reduce the need for resources and effort to grow food we know will be wasted.

When a business uses the word vertical they usually refer to the vertical supply chain from originator, down through distributors and agents to retailers and then down to the final consumer. When talking of complex items, such as cars with many component suppliers, the chain becomes very complicated. Originally manufacturers made everything themselves. By the 20th century it was normal for many components to be bought in as mass production enabled companies to specialise in individual components. As often happens practices that started as a cost saving measure see the pendulum swing to the opposite extreme and the cost saving principle is lost. Some then attempt to gain control of the chain with vertical integration – often buying or finding a way to control the supplier next up the chain. Yet this is the one bit that is rarely, if ever, considered by those setting up vertical farms. To make economic sense vertical farming should mean growing on the retailer's roof and this was the conclusion of a report commissioned by the UN. When looking at costs this is a fundamental change and avoids all costs of distribution but does need a new business model.

2.2 Costs of Distribution

Many people talk of the need to reduce 'food miles' – the distance food has travelled. The most obvious point about distribution is the one that too many urban farmers miss completely yet is taught to teenagers. The final mile of distribution is always the most expensive. Modern container ships talk of achieving over 1000 ton miles per gallon of diesel. A modern truck carrying a shipping container can achieve 100 ton miles per gallon on a trunk road. On a city road this will halve. Worst of all will be while waiting at a junction for traffic lights to change where this figure will be less than zero – at a point where town councils are now siting monitoring equipment to measure exhaust pollution.

These figures will be for average weight containers. When shipping food products there is big difference between the bulk movement of grain and lettuce. Grain can be shipped without the need for packaging. Retail packs of potatoes can be shipped in bags stacked on pallets making up a few percent of total shipped weight. Iceberg lettuces need to be packed in strong cardboard boxes on pallets getting near 50% of total shipped weight (a 40' reefer shipping container can carry 20 pallets x 85 boxes x 12 pieces = 20,400 iceberg lettuces). Then there are items like fresh herbs which require far more packaging which can end up near 75% of total shipped weight much of which is only necessary to prevent crushing in transit. This means many salad and herb varieties cost traditional growers more to deliver than to grow (even though many urban farmers forget to include the packaging as part of the costs of distribution). Simply reducing food miles therefore achieves very little. Delicate crops need the transit only packaging to survive being placed on the delivery vehicle no matter how short the distance.

When looking at the carbon footprint of food there are several reports that conclude that, when comparing calories, the carbon footprint for lettuce is several thymes that for bacon. Plawecki, Pirog, Montri and Hamm wrote a paper [9] in 2013 giving a comparative ratio of 4.3 to 1. In their effort to debunk this Grist.org [10] said it was not correct to compare calories before admitting it would take 93 times as much lettuce to give the same calorie count as bacon. Whatever one's views on the accuracy of such thinking the author argues that is cannot be right to swap a poor process for a terrible one.

In June 2108 Eric L Adams, Brooklyn Borough President opened the Indoor Agtech [11] event in New York where he spoke of using diet and exercise to reverse type 2 diabetes and the deleterious effects he had started to suffer. He also spoke of how it was the low paid who most needed to improve their diets yet were least able to afford to do so. New York's policy of helping the poor is not like the UK's use of 'foodbanks' but to offer those who qualify vouchers that can be exchanged for fresh produce. This scheme has now gone beyond that shown in the presentation by Catherine Luu [12] titled 'Pharmacy to Farm' which has a focus on those taking medication for high blood pressure and now includes 'Heath Bucks' to increase the value of goods bought from local farmers' markets [13]. So, whilst there is a need to increase the amount of food that reaches the consumer, it must be remembered that spending vast sums of money to grow more may provide micro herbs as garnishes for top end restaurants but it will do little, if anything, to help feed a busy city. There are also many other reasons we should discourage the growing of micro-greens as highlighted in the later section about seeds.

The bottom line for distribution is that simply reducing food miles does little to address the need practitioners suggest they are addressing. Removing the need for the final mile can achieve far more.

2.3 Traceability

The need to know where food originated has never been greater. The more links along a supply chain the more opportunity there is to lose track of where the chain started. There are several recent examples of how important this has become; both to prevent fraud or to know the origin of food poisoning.

Manuka honey is a product highly valued for its heightened antibacterial and anti-inflammatory qualities which can be many time greater than other honeys [14]. It is created in New Zealand by bees with access to manuka pollen. The premium prices that can be achieved have attracted the attention of fraudsters with current efforts to trace the source of origin all the way through the supply chain failing. The scale of the fraud is known because the claimed sales are significantly greater than the quantity that can be produced by the original source.

This fraud is obvious but far less dangerous than food poisoning. This was made very plain to see in the summer of 2018 in the USA where five people died and 200 fell ill from E. coli spread across 35 states. The source was quickly narrowed down to romaine lettuce but it took weeks to narrow down to the source farm and longer to find the field. The complexity of the supply chain is shown in Figure 1. Despite the scale of the problem this is only one of many recent outbreaks and an ongoing problem throughout North America. In December 2018 the USFDA issued a major recall of all romaine and several other varieties [15]. The recall meant there was no availability of romaine lettuce regardless of where it was grown. One of the big selling points of growing in protected facilities is the ability to shut out pest and disease yet even these growers were not allowed to sell their expensively grown produce. The public concern caused by the summer deaths meant that perfect produce that could clearly be seen to have come from a clean source was wasted. The result was the public not getting the food they wanted, many farmers lost valuable income and the average environmental costs for delivered produce increased.

In the UK retailers are required to be able to give the source of origin for any food product within four hours. Whilst this is simple for products such as tinned goods where information has been printed on each tin this becomes much more complex for loose goods and produce where information is collected and recorded manually. This is where AI can really make a difference. If the data is there it is possible to give the source of origin simply by scanning the same barcode as used by the retailer.

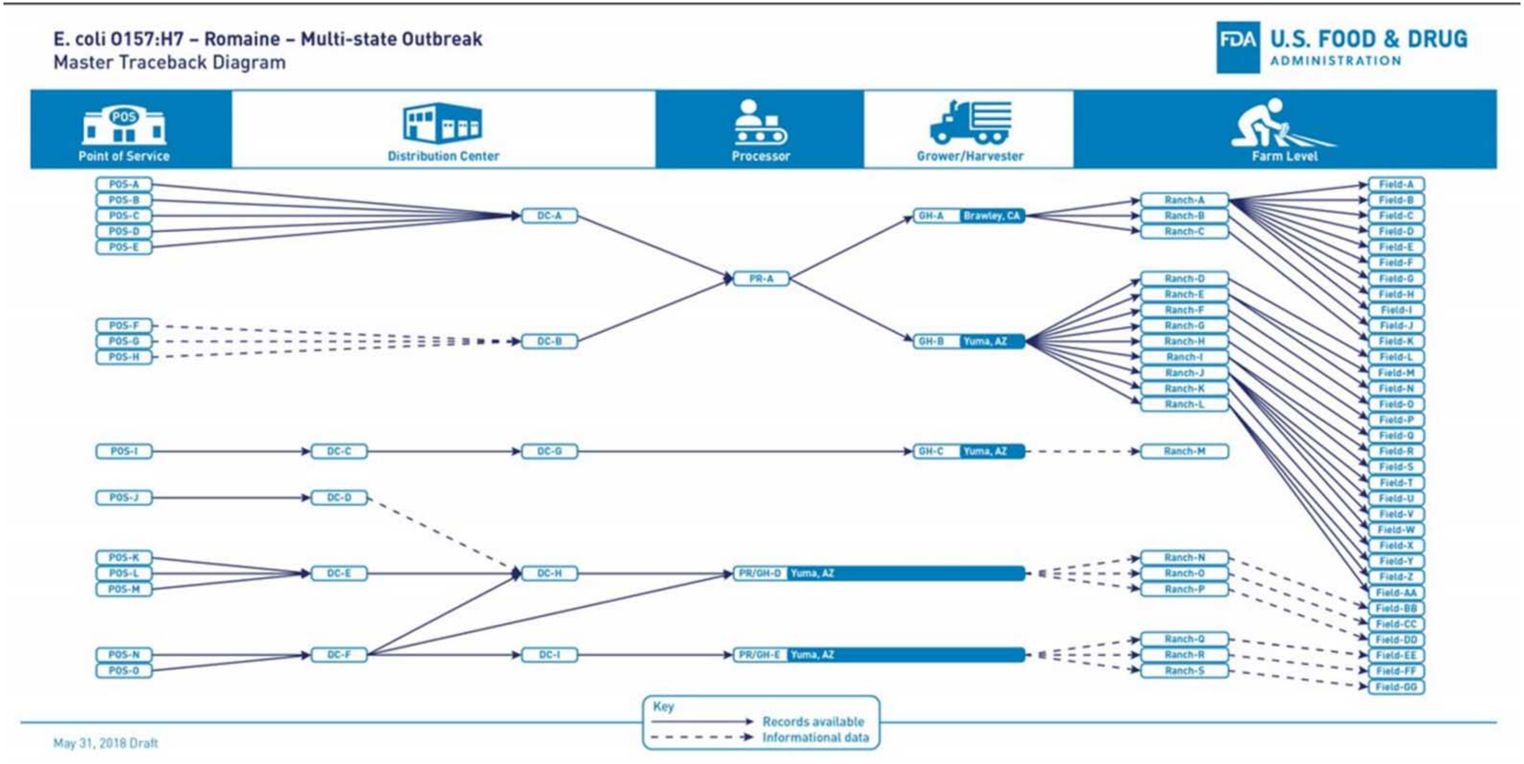


Figure 1. USFDA shows the complexity of their fresh produce supply chains. 2018 [15]

2.4 Distribution

Big food processing companies such as Mars Inc. and Unilever refer to their products as Fast Moving Consumer Goods (FMCG). Invariably this refers to goods that are sold in large numbers produced in large scale yet having a long shelf life. They are produced at high scale and delivered via whole truck loads to warehouses where they can be stored for several months. Despite this definition the fastest moving consumer item is a daily evening newspaper. They are planned and printed in the morning, need to be on the newsstands by lunchtime and are waste by late evening. When it comes to agriculture the most commonly used meaning for FMCG fits with arable farming and a similar set of circumstances. Most cereal grains are harvested once a year and therefore, by definition, are stored on average for almost half a year. Despite this, or maybe because of the qualities needed for storage, the science applied to these crops is often greater than for short cycle crops such as salad leaves which fits the example of an evening newspaper.

The pluralisation of the word ‘cost’ is deliberate. Monetary cost is only one part. We are now increasingly aware of the environmental costs of supply chains. For fresh produce this is far in excess of many other sectors. A UK bed manufacturer was once quoted in a trade magazine as suggesting they specialised in packaging and selling fresh air. Most distributors go to great lengths to avoid the need to move large volumes of fresh air as it costs so much for little, if any, profit. This is taken a step further by the wine industry where even mid-priced wine is exported in bulk tankers and bottled in the destination country.

When it comes to delivering perishable produce salad leaves and fresh herbs rely on disproportionately large amounts of single use transit only packaging to avoid the produce being harmed. Most fresh salad crops are easily crushed and can dry out rapidly once harvested which largely defines their shelf life. Some talk of the nutritional value of a lettuce reducing to 50% within 2 days of harvest yet some take longer than that to reach the retail shelf.

Most teenagers learn that distributors and delivery companies talk of the final mile as being the most expensive. Some retailers take this a stage further and talk of the last 50 metres being the most expensive. This is more about the logistical tasks of unloading the truck, checking and signing for the delivery, temporary storage and only then the task of moving and loading the produce onto shelves. Even this is not the end with the need to store reusable delivery trollies and then to find a responsible way to dispose of single use transit only packaging. Thinking about this should suggest that moving the traditional commercial greenhouse model to an expensive city building makes little sense. Only in a few cases of extreme distance can saving food miles offset the increased overheads of a city building. Looking further into the economics and productivity of what many now call vertical farming one can see many reasons to avoid reserving a whole city building for the growing of food.

2.5 Urban farming

Having looked at several aspects of CEA it can be seen there is a need to look more carefully at what it means to be sustainable. With the extreme costs of reserving a city building to growing food the use of traditional business models makes little sense. There are many examples of investors losing their money as investments fail. Indeed, whilst aiming to publicise a recent investment in the UK the Financial Times journalist Leila

Abboud mentioned that Plenty Inc in America had suspended development of their second facility whilst they researched how to avoid making such heavy operating losses in their first [16]. It is very strange that investors are simply not undertaking adequate due diligence when they make their investments in some sectors.

Far too many who set up in enclosed spaces seem to be relying on what could essentially be referred to as adding complex plumbing and expensive lighting to warehouse shelving our grandparents would recognise.

There are some who take advantage of the low prices and ready availability of used shipping containers. Depending on manual handling, installing rigid shelving and reserving a third of the floor area for an access corridor does not make effective use of the space. Those who fit out these containers make claims about productivity that few of their customers get close to achieving. Some make claims these facilities can grow as much as one, two or even five acres of traditionally farmed land. The fact that there is an active market for second user container farms only a couple of years old shows they are not making the suggested return on investment. This is not at all surprising when seeing the actual quantities harvested. They are often promoted as scalable. They certainly are for those selling them but not for those buying them who find they are only repeatable. If you need to get a coachload of people across the country you don't hire a 10 strong fleet of taxis with 10 drivers, 10 engines and control units, 40 wheels and 10 aircon units when the same can be achieved with a single coach and driver. So why would you choose to farm using a fleet of taxis?

There has to be a better way. At City Farm Systems we believe we have found a solution that relies on data to address many of the difficulties others can't. In 2017 the author was invited to join a trip to China. On his first full day there he was asked to extend his stay so he could deliver a keynote speech at an international conference. With little time to spare he had to prepare a presentation for a new audience and give it a title. The conference was chaired by the Dean of China Agriculture University's College of Information and Electrical Engineering, Professor Wanlin Gao [17]. Using this as a starting point the presentation was titled 'Using Data to Unlock the Potential that can only be found by Growing at the Point of Need'. This talk followed on from that given by Professor Nick Sigrimis [18] from the Agricultural University of Athens. During his talk Professor Sigrimis spoke about the need for urban agriculture to do much more to collect and use data as it was falling behind traditional agriculture. He also suggested that the use of RFID would enable this. Professor Sigrimis finished by suggesting there was little to be learned from the urban farming sector in America. Following directly after this the author started by agreeing with those thoughts and then set about showing how data could make a difference.

Perhaps the best reason to look at this is that there many reasons fresh produce should not be made more expensive. As can be seen with the USFDA chart (Figure 1) showing the romaine lettuce supply chain it is very complex with more costs added at every step. Urban farming was started as a way to increase food production and suggested a reduction of this complexity yet relies on the same business model. It is not the distance that creates the cost but the number of links along the supply chain where each transaction adds a price uplift as each operator seeks a profit defined by percentages rather than transactional costs.

There are some urban farming operators who try to add the data collection and usage that is improving some areas of traditional agriculture but the complexity of their systems makes this all but impossible. Plenty Inc. in America talk of using 7500 infra-red cameras and 35,000 environmental sensors in a 55,000 sqft (5116m^2) factory farm and maybe this is why they have suspended the development of their second facility [19]. That is almost 1.5 cameras for each square metre. The complexity of powering and connecting that number is incredible. Wireless cameras still need power connections. At the same time the number of images needed to give a time lapse video of each plant is at most a few a day. This means the cost of each image collected is enormous. Each camera is capable of delivering many images a second yet is only needed to save one or two per hour at most. This is a typical issue for far too many expensive items used in agriculture. It is not hard to imagine the conversations that would take place in finance departments if the combine harvester was a new invention. At the UK's Institution of Agricultural Engineers' 2017 conference Sean Lennon spoke of the average modern combine harvester working for 18 days a year [20]. That's only a 5% usage rate if it was working 24 hours per day and needs an enormous amount of grain transportation and handling kit available to keep it working. It also needs to be started regularly throughout the rest of the year to ensure the expensive computerised control systems are maintained and still capable of running when needed the following year.

With CEA there are many examples where a little thought could have reduced the amount of expensive technology required allowing the investment to be used in areas that solve other issues. Moving crop trays around a naturally lit greenhouse can be achieved with simple, low stressed mechanical elements avoiding the need for areas devoted to occasional access. This means a greenhouse can be installed in a location where the problems of insurances and Health & Safety would make regular workforce access prohibitively expensive.

2.6 Sustainability: is it correct to claim CEA as sustainable?

Most CEA practioners talk of it as a sustainable method of feeding a large population. The reality is, however, that the majority should not be making this claim. When talking in business terms the most fundamental aspect of sustainability must be the requirement to make long term profits and making a positive return on their investment. The recent promotion of second hand shipping container 'farms', the number of companies closing down and the number of heavily invested companies not expected to make a profit for a considerable time suggest they had not addressed this basic business requirement.

Then there the claims about the virtual increase in area of cultivated land. This can quickly be discounted when looking at the use of energy where, if they were to use renewable electricity generated by Solar PV panels, the area of land needing to be covered with panels would by far in excess of the land they claim not to need. Very good Solar PV panels are only near 25% efficient…when directly facing the sun. In the UK where the PV industry talks of one hour of peak sun in winter rising to five hours in summer. The figure quoted by one of the agencies that helped Plenty Inc. raise their $200m is that five times the growing area must be covered with PV panels although is an average and needs to be considerably great for winter sun.

Then there is the issue of their seed requirement. When growing herbs rosemary is an exception (usually propagated by taking cuttings rather than growing from seed). Seed supply is a far more important point issue than many realise and can be problematic – especially for crops where the desired harvest is not normally their seeds. Take a look at the small number of plant families we rely on. The brassica, melon and allium families give us an enormous range of food crops. Not many realise that oil seed rape is a brassica grown for its seeds. Then look at Brussel sprouts and cabbages. They are both brassicas that have been bred for very different characteristics where flowering is an unwanted characteristic.

Celery is a crop that is very problematic when it comes to seed. For every 10 acres of crops grown for harvest there needs to be an acre devoted to growing plants for seeds. Food retailers have responded to consumer demand for smaller versions of many varieties. If one follows this trend, then the size for a celery plant would be roughly one third of the height of those more normally available. Using basic mathematics, it can be seen that to grow the same volume would require 27 times as many seeds. So now the area of land devoted to growing the seeds is not one in 10 but 2.7 times as much as the food growing area. To be truly sustainable a CEA facility needs to address all of these issues.

We are now increasingly aware of the need for a sustainable supply of healthy nutritional and affordable food. However, it can be seen that much about CEA is not sustainable but Artificial Intelligence can go a long way to help change this. Looking back at the comments made by Eric L Adams we should also pay more attention to costs. Real estate agents talk of the value a house can be expected to reach never reaching significantly more than the locality will bear. No matter how much you extend a house in an average street it will not command a significantly higher price than those around it. The same applies to growing food. There will only ever be a few who are prepared to pay significantly more than the average price. To feed a city we need to look at affordable food. The last thing we should do is to increase costs above that which the market will bear. For CEA this has to mean finding ways to avoid costs rather than to add costs that do not improve the produce.

3. Growing Methods: A Review of CEA Systems

There are several distinct approaches to growing in a closed environment. As discussed they are all variations of hydroponics when one considers that this is the way in which all roots absorb water and nutrients. There are, however, several ways in which the delivery of nutrient rich water is delivered. This section shows the most relevant ones.

3.1 Hydroponics

One of the first questions asked of any new CEA project is ‘do you use hydroponics?’ This is an interesting question. The term is widely used to refer to the growing of plants without conventional soil and where the nutrients needed by a plant are delivered by salts dissolved in water. However, when studying plant biology, it can be seen that this is how all rooted plants take in their nutrients. To understand this the needs of a plant must be understood clearly. A plant’s roots achieve two key needs. They anchor the plant and deliver water and nutrients needed for the plant to grow. Those who use Hydroponics often talk of soilless culture. No matter how a plant is grown roots can only take in nutrients dissolved in water. Those who have studied chemistry will know a dissolved salt is no longer a single compound as it divides into two ‘ions’; one with a positive charge and the other with a negative charge. The important difference is the way in which the supply of nutrients is buffered. Using water alone to deliver nutrients means those the plant needs must be present when needed. This becomes one part of why closed environment facilities that depend on whole system watering systems are restricted to growing a narrow range of crops. When grown in soil nutrients are either ever present or added during the crop lifecycle. In this case the changing needs of the plant and any variation of nutrient uptake must be triggered by something such as change in temperature or light quality rather than simple availability. This is a very interesting topic and it can be shown, just like animals, plants react to a circadian rhythm. [3] In their study TiMet go as far to show that some plants have an internal clock that counts down the time until they see sunlight each day. This directly discounts the claims some 'factory farms' make about being able to power their LED lights with lower night time energy costs and to switch lights on or off according to buffer other's energy demands. Specifically, '*...the TiMet team was able to show how starch regulation maintains the plant's optimum growth rate in a 24-hour rhythm. Moved to a 20-hour cycle in a laboratory (with ten hours of light and ten of darkness per cycle), every artificial dawn surprises the plants and their growth rate significantly decreases.'*

Another point that must be considered is that roots also need oxygen to remain healthy. So for a successful Hydroponic system the roots must have access to air or the water needs to be oxygen enriched (see Figure 2). This is where the need to anchor the plant becomes important. A healthy soil needs to be aerated as well as holding water.

Hydroponics splits into three distinct types. One uses inert media with large air gaps held in a pot, the second puts young plants into a plastic cage that fits into a plate that forms a lid above a closed volume of water or the third version where the cages are placed into slabs of polystyrene floating on a vast tank of water and usually referred to as floating rafts.

In the first case it is common practice for every plant to need an individual water supply that ensures the growing media is kept moist. There is an alternative to this where each pot or a set of pots is contained in a bigger container where water can be pumped in on a regular basis before being allowed to drain away. This is commonly referred to as 'flood and drain' or 'ebb and flow' (see Figure 3). This allows a whole system to be watered thoroughly and, as water is added from the bottom the substrate is not compacted and as it drains away air is drawn down and ensures the roots get the oxygen they need. This is a very effective way of encouraging strong root growth as they are encouraged to grow down to the source of water. Such a system needs thought about how to make the drainage work effectively. The more common methods require a lot of complex pipework (see Figure 4) to enable the manual insertion of small water delivery tubes into each growing container and thus relies on large workforce access areas reserved for occasional use. Invariably all these systems rely on a facility wide supply of nutrient rich water monitored by EC sensors and complex pipework.

Having noted the points above then the answer to the question 'do you use hydroponics?' should be that all plants grow hydroponically. All plants with roots depend on taking in a range of chemical elements in ionic form. What needs to be considered far more is that soils have trace elements that are not yet monitored in any great detail. There are trace elements and minerals that naturally grown plants absorb. The list of elements animals and humans need is greater than those needed for healthy plant growth. And this is why we should look further at what this means for agriculture as a whole with quantifying these elements being the first stage and using AI to understand how and what should be done to optimise human nutrition.

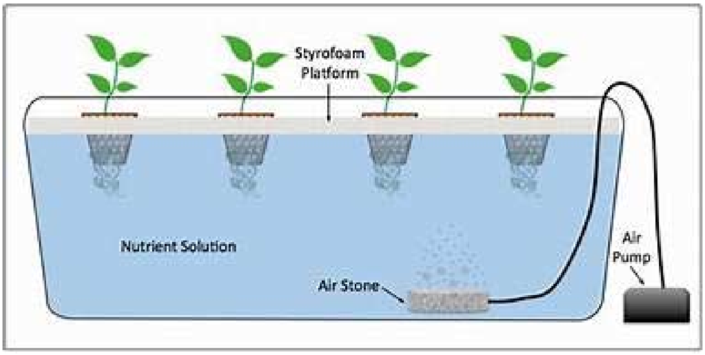


Figure 2. Floating raft hydroponics

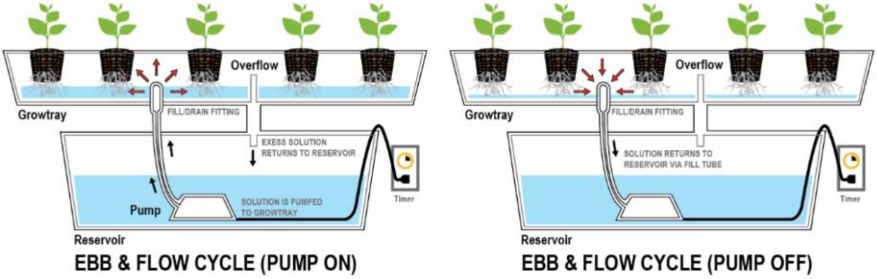

Overflow
Growtray
FILL/DRAIN FITTING
EXESS SOLUTION RETURNS TO RESERVOIR
Pump
SOLUTION IS PUMPED TO GROWTRAY
Timer
Reservoir
EBB & FLOW CYCLE (PUMP ON)
Overflow
Growtray
FILL/DRAIN FITTING
SOLUTION RETURNS TO RESERVOIR VIA FILL TUBE
Timer
Reservoir
EBB & FLOW CYCLE (PUMP OFF)

Figure 3: Flood and Drain – avoids complex pipework

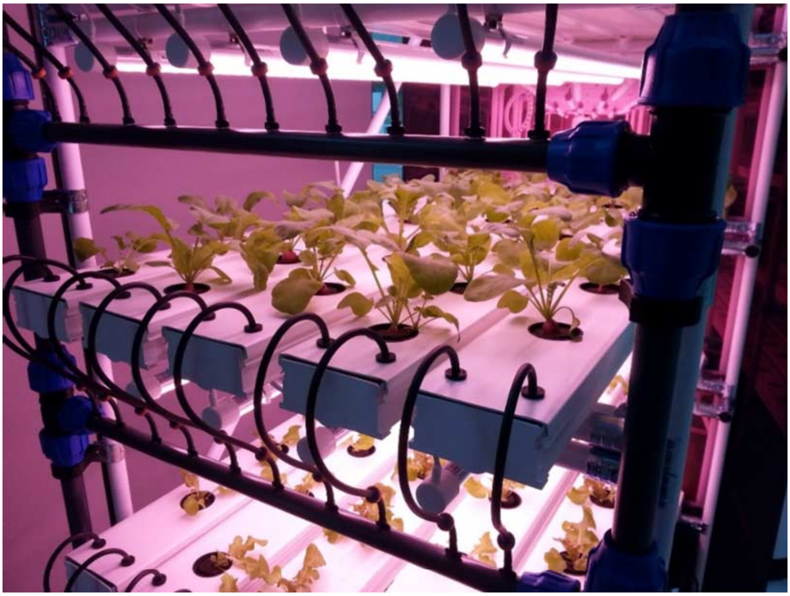


Figure 4: Complex pipework makes more sense for whole season vine crops than short cycle crops.

3.2 Aeroponics

This is a variation of Hydroponics where the plants are held in such a way their roots can be in a near constant mist spray of nutrient rich water. This has been promoted by several as being more effective and requiring less water than hydroponics (see Figure 5). This can be disputed however as it is only the amount of water held within the system that changes and there are other downsides. Until recently the majority of spraying nozzles used to create the water mist needed extensive maintenance to avoid clogging as the nutrient salts crystallise. Nozzle design has improved and this can be less of a problem with some being developed at great expense despite the whole process failing to address other issues.

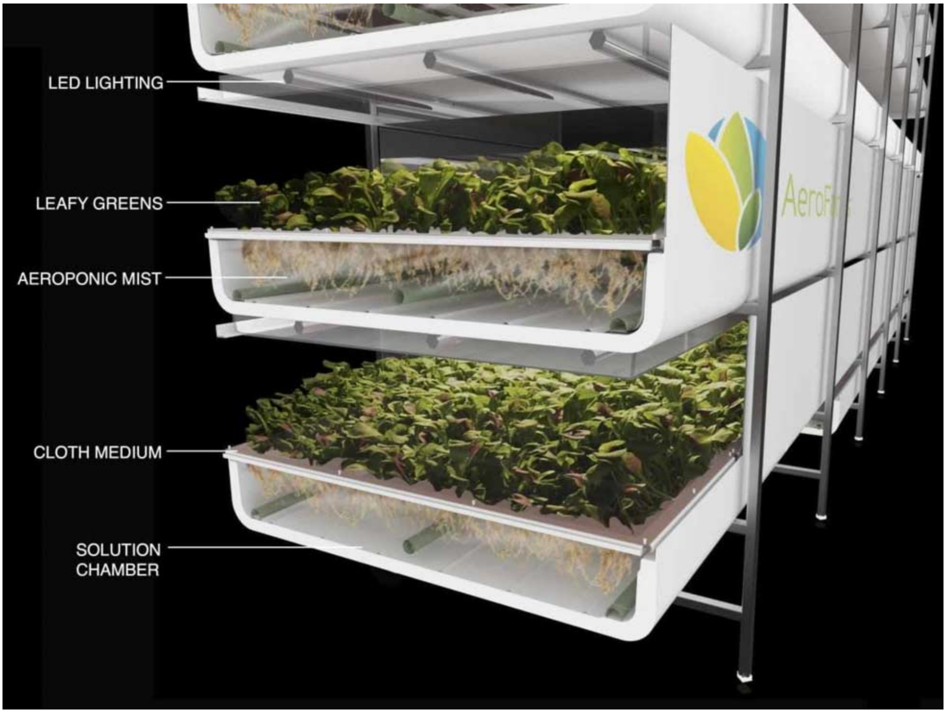


Figure 5. Aeroponics as used by AeroFarms in USA. [21]

Plant biologists know how complex a leaf can be. In some ways they can be thought of as complex as human skin (see Figure 6). Growing in a tightly controlled facility that sets out to maximise the rate of growth has some unintended consequences. The underside of leaves is where guard cells control stomata which are pores that allow carbon dioxide to enter the plant and unwanted oxygen to return to the air along with moisture if in excess or to cool the plant. This is one of the most important aspects of any circular ecology as it is the opposite of mammals breathing in oxygen and exhaling carbon dioxide and which leads to the part of global warming caused by deforestation.

In an aeroponic system the concept is about avoiding biotic and abiotic stress. When there is never a shortage of water the guard cells have a much reduced need to close their stomata. As a result of some suggest they lose the ability to close although others argue this is a fundamental part of the response to a circadian rhythm. It is now understood by some that aeroponically grown crops are likely to have a shortened shelf life. Normally, when a lettuce is harvested, stomata are quick to close and hence reduce the transpiration. When stomata lose this response they remain open after harvest and the crop dries out quickly and defeats one of the main reasons for using this growing method. This would not be a problem when growing strawberries or other crops that are not grown for their leaves. This also suggests that the floating raft version of Hydroponics suffers from the same problem. The floating raft method also adds the need to ensure the water has a sufficient oxygenation level. Fortunately, the nature of floating rafts largely prevent light reaching the water. This needs to be ensured to reduce the level of algae and mould growth which consume nutrients and require cleaning and/or removal.

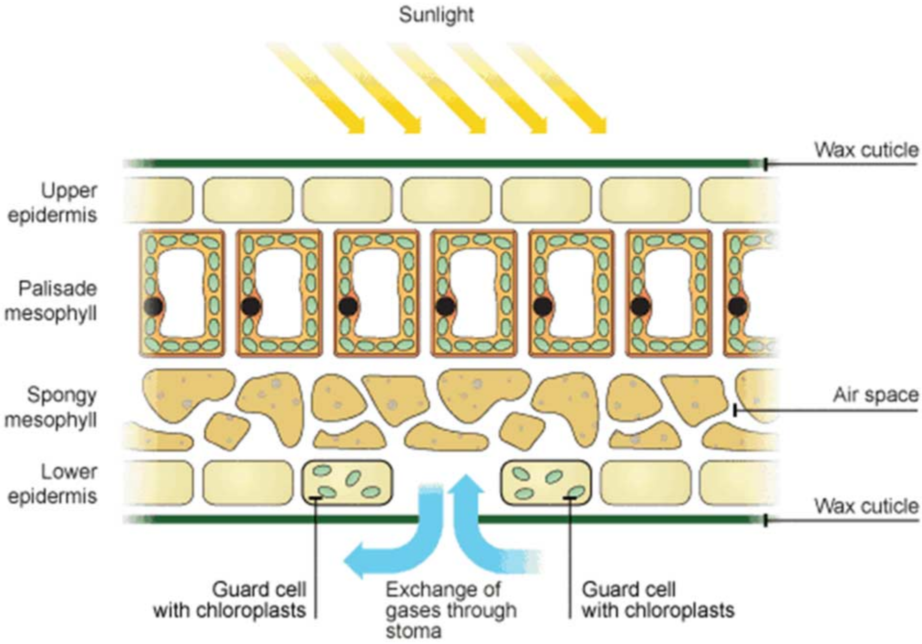


Figure 6. Cross section of a typical leaf showing the structure

3.3 Aquaponics

Aquaponics has been hailed by some such as the team at AquaGrove in America as the ultimate circular farming activity combining fish farming with hydroponics [22]. Some hail it as a new concept yet the concept can be traced back thousands of years as a standard part of growing rice in paddy fields. More recently the possibilities were seen in the UK when flooding enabled fish from a trout fishery to enter watercress beds. The result was the watercress grew rapidly and the fish were healthy despite not having access to their normal sources of food. At these lower stressed levels, the concept makes sense.

When transferred to intensive indoor facilities the limits are found quickly. Several aquaponic farms in America such as Farmed Here have shut down with at least one suggesting that the fish element added 30% to their overheads. As the hardest part required for success it is also easiest to avoid. Invariably the most successful fish grown in indoor facilities has proved to be tilapia; a warm freshwater fish happy to live in high densities. It would seem that unless there is a very strong local demand for these fish there is little point in attempting this highly stressed way to grow food. It has been suggested that the best use for tilapia is to process them into feed pellets for salmon farming – which means aquaponics is not a viable business model for urban farming. It would seem that using aquaponics to add salad growing as a secondary income for a fish farmer could make sense but that adding fish production to the production of salad crops is a very expensive way to restrict what can be grown (see Figure 7).

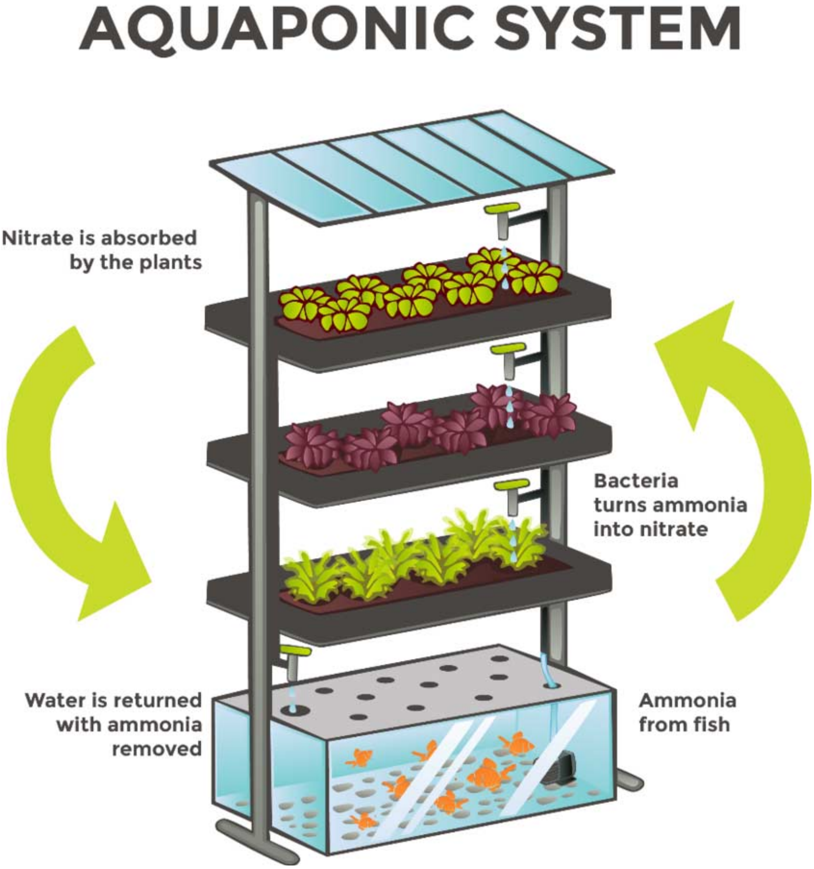

AQUAPONIC SYSTEM
Nitrate is absorbed by the plants
Bacteria turns ammonia into nitrate
Water is returned with ammonia removed
Ammonia from fish

Figure 7: Aquaponic cycle [23]

3.4 Artificial Lighting

Lighting is a fascinating subject for agriculture and far more complex than is obvious. It has been suggested that the idea of using blue and red LEDs came from the Phillips when the blue LED was first created in 1994. The colour of light coming from a light emitting diode is defined by the voltage used to excite the chip. This is a result of the atoms involved with electrons circling the nucleus at differing orbits. The higher the orbit the greater voltage needed to create the emitted light and each orbit offering a different light wavelength.

The energy coming from the sun is a spectrum that is wider than we can see. The spectrum we see starts with shorter wavelength blues and extending through the rainbow spectrum up to longer wavelengths of red. Wavelengths shorter than blue are referred to as ultra-violet. Ultraviolet light in the C spectrum (UV-C) is energy-rich light with a wavelength of 200 – 400 nanometres (nm) [Figure 8]. UV-C light is very versatile and can be used for disinfecting water, destroying harmful micro-organisms in other liquids, on surfaces, on food products and in 'air'. With UV-C technology it is possible to kill pathogens with more than 99.99% destroyed in seconds. This avoids the need for chemicals and has no long-term harmful side effects provided operators are protected. UV treatment is inexpensive, highly efficient and absolutely reliable. UV is also widely connected with the anti-oxidant properties of a plant.

Infra-red light has a longer wavelength and has a big influence on a plant’s colour. Controlling this can have interesting implications for variegated leaves. LED manufacturers talk of light recipes as the particular emphasis on parts of the colour spectrum. This is an interesting point where the way a plant reacts is not solely down to the presence or lack of a particular quality of light. A lawn specialist will talk of the strength of green being dependent on the soil nitrogen level. Basil growers will talk of the qualities gained by adding red light late in the crop cycle. So the desired quality is not created simply by changing light or nutrient level. However, nature suggests light would be the key driver or trigger.

It should be noted that sunlight has a natural variation in colour levels during the day. We have red sunsets and sunrises because the shorter blue wavelengths are reduced as sunlight has to travel further through the earth’s atmosphere. The shorter wavelengths bounce around the air molecules more than longer wavelengths and so have more difficulty getting through and is why a clear sky appears blue. This not only means UV levels are higher at midday but that the further from the equator we are the greater reduction we see for blue light in winter.

There is a need to look at light in ways we humans cannot. We need to consider parts of the sun’s spectrum beyond that which we see. A large part of this is that the human eye has dedicated green light receptors unlike other animals. Reindeer, for instance, can see infra-red which is how they find lichen in snow covered ground. It has been found that adding UV light to an intensive chicken rearing shed will reduce the amount of fighting and broken bones amongst birds. We humans need to see green light; we have dedicated green receptors in our eyes which means we see differently to many other animals. The idea that plants don’t require green light is therefore a very simplified thought that has very good reasons to be ignored. Specialists that study this can show that plants need yellow and green light to control certain growth characteristics.

Osram is one of the two original patent holders for the manufacture of light emitting diodes (LEDs). The overwhelming majority of LED ‘manufacturers’ buy the actual diodes from Osram or Phillips and then add other electronics to power or ‘drive’ them.

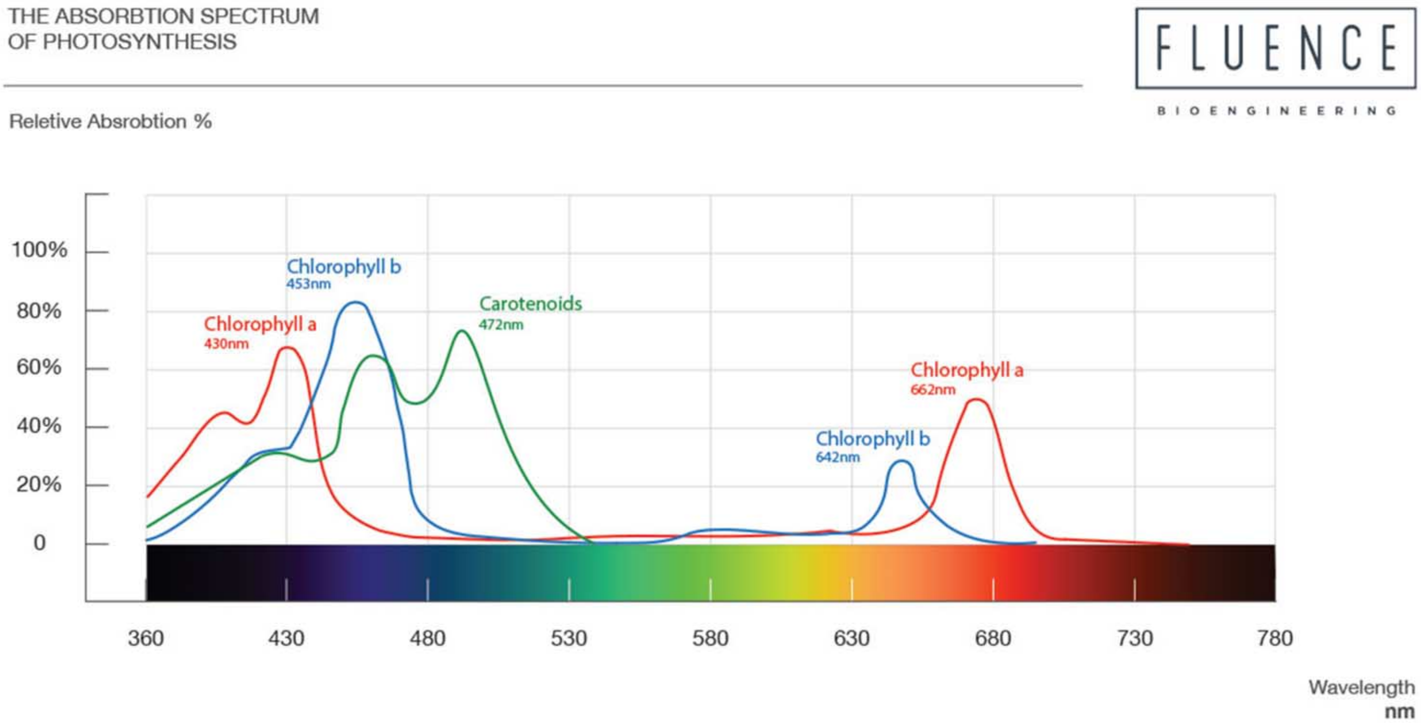


Figure 8. The absorption spectrum of Chlorophyll a and b measured in vitro. [24]

This quote from Osram's Fluence Bioengineering team speaks volumes: *Many manufacturers reference the absorption spectrum of chlorophyll a and b (which peaks in the blue and red regions of the electromagnetic spectrum) as the main reason for providing a purple spectrum* (Figure 8).

At first glance, this seems legitimate since chlorophyll drives photosynthesis after all, but have you considered how the absorption spectrum of chlorophyll is measured? Additionally, have you considered if the absorption spectrum of chlorophyll directly correlates to photosynthesis and plant growth, and what happens if you only target a sole pigment and neglect other pigments responsible for plant growth and development? This article will discuss the differences between absorption spectrum and action spectrum, and (spoiler alert) dispel the myth that "plants don't utilize green light" to promote plant growth and development.

This only starts to show the importance of light quality when looking at artificially lit growing facilities. Invariably they light from above but it can be seen that in nature a plant does not experience midday light for the whole day. Plants react to changing light qualities. Perhaps the most obvious can be seen in a field of sunflowers. At sunrise the flower heads will be facing east and as the sun tracks across the sky so the flowers turn to follow it and ends the day facing west. Thinking about this one can understand a lot more about how a plant develops. Light from the side will be stronger in the longer red wavelengths and therefore contributes to spring and autumn (fall) qualities. Obviously in autumn this has a bearing on the ripening of fruit.

Plants are incredible things. Not only do some track sunlight but others, like some bean varieties, do not flower until they have witnessed mid-summers day and start flowering when they identify that day length is reducing. Others react to temperature whether that be the need for a cool night or suspending growth when the average temperature across the daylength is too high. All of this means that the subject of lighting is very complicated. Using a fixed light recipe in a confined space severely restricts the plant qualities and varieties that can be grown. To understand this fully means the need to measure different aspects. Plant breeding is a very interesting topic when considering these points.

The author was again at China Agriculture University in Beijing in late 2018. Whilst there he had a long conversation with a visiting professor from Pakistan. Professor Mateen Khattak of Peshawar Agriculture University [25] studied for his PhD at Reading University in the UK where his work was based on ornamentals rather than food crops but showed that adding chemicals to polytunnel covering plastics could be used to selectively filter sunlight. This was before the ready availability of LED lighting. The results are now integrated into major polytunnel plastic film manufacturing [26].

Figure 9 shows a clear film with a slight variation in UV light but is otherwise constant across the spectrum and recommended for simple plant protection. Figure 10 shows mostly blue light can be let through and is used to slow the growth of ornamental nursery stock ready for sale and Figure 11shows all but green light can be restricted and is used for shade loving plants as if to replicate a forest floor position.

The expense of artificial lighting has come down significantly but is still very expensive to install and power. This is a point which many CEA specialists fail to address. They claim LED is energy efficient and, of course, in comparison to older generation lights they have made significant improvements.

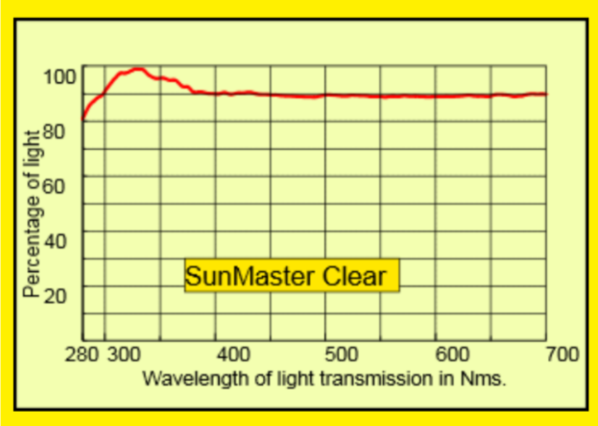


Figure 9. Altered UV light Filtering Qualities of Plastic Films [22]

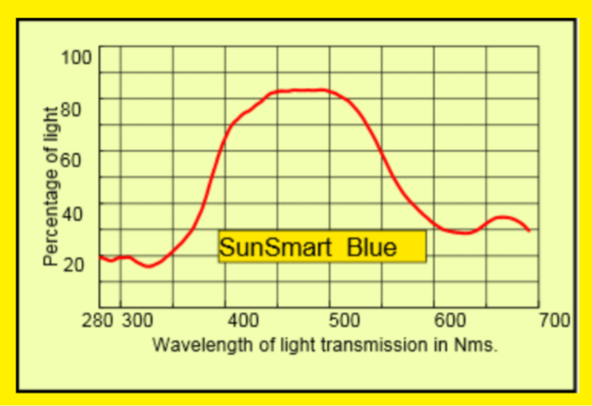


Figure 10. Altered Light Filtering Qualities of Plastic Films [22]

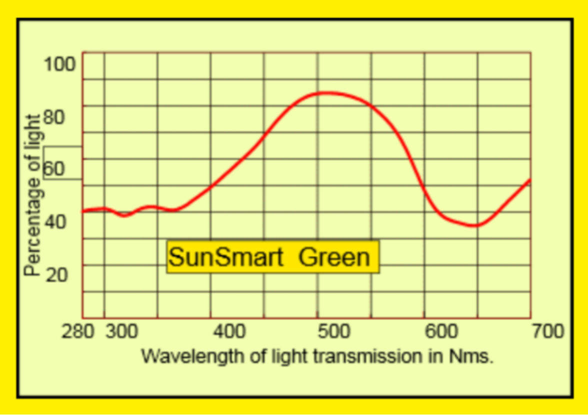


Figure 11. All but Green Light Filtered from Plastic Film [22]

Most practitioners make huge claims about sustainability but what they don't do is admit just how much power they consume. Some will say they have contracts that ensure the power they consume comes from renewable sources. What they do not admit is that using solar PV to power their lights requires a greater area of land to be covered in panels than the amount of land they claim to save by growing more intensively. Looking at this further it can be seen that, at a time of immense pressure on power generation, reserving the use of renewable power for horticultural operations that would do better to avoid it prevents other more urgent uses. This includes the charging of electric vehicles used for taxis and buses in heavily congested cities.

There are also other significant downsides of artificial lighting. The author has visited sites and seen groups of visitors feeling noticeably uncomfortable within minutes of being subjected to blue/red LED lighting. He has also seen shops where the lighting was painful to experience. The human eye needs to see a light spectrum that includes green. This is something that is very noticeable in city buildings. A small amount of natural light is enough to make artificial lighting much less unpleasant. This approach has been seen to have benefits for many different animal species. There should be no surprise there is also evidence to show plants need to experience natural sunlight. At the same time just as we shouldn't stay out in intense summer sun for too long the same must apply to plants. So for a healthy plant any intensive growing facility should use naturally varying sunlight in preference to constant artificial lighting wherever possible. Professor Qichang Yang at the Chinese Academy of Agricultural Science (CAAS) in Beijing [27] has carried out intensive research into the ideal planting density of crops in natural light. So why pay to shut out sunlight and then pay even more in a vain attempt to recreate it? There is far more to be achieved by adding supplemental lighting for short periods of time. In a system where crop trays are moved constantly these lit areas need only be a small part of the whole facility and therefore offer much reduced capital expenditure and running costs.

3.5 Carbon Dioxide – CO_2

CO_2 obviously plays a fundamental part in plant growth – it is one of the primary requirements much like a fire needs oxygen, heat and fuel. When climate change first started hitting the headlines ambient levels were around 350ppm. In an intensive greenhouse facility if CO_2 is not added growth will stop regardless of the other needs. Air must be brought in from outside or CO_2 introduced in other ways. It is interesting to note the varying levels growers use. Some talk of maintaining 600ppm, others exceed 1000ppm whilst one American grower suggested lettuce growth is maximised at 17-1800ppm. Traditionally greenhouse growers have used heaters to encourage growth. Those with gas boilers or even gas powered combined heat and power units (CHP) have been able to vent their flues and exhausts directly into the greenhouse.

Crops in very hot climates crops can also stop growing. Factory farms are able to reduce temperatures but in those areas where this makes most sense such as some Arab states they are not permitted to add CO_2. There are instances where the growing environment is cooled far below the ideal growing temperature so that achievable growth rates match the available CO_2 levels and the speed at which external air can be cooled and introduced.

There is still some work to be done to understand how varying levels impact on growth along the growth cycle. Sweet pepper growers talk of young plants suffering 'growing pains' if levels are too high. This can even reduce the quality and, therefore, the value of the fruit with visible scarring.

Measuring CO_2 levels is essential but requires sensors that have a heated element. This makes wireless sensors problematic as power requirements are much higher than other sensors. The use of AI would enable the CO_2 requirements to be much better understood. This would be particularly beneficial along a crop's lifecycle.

4. Nutrients: why Measuring Nutrient Level is Critical

4.1 Nutrients for Plant Growth

Nutrients can be a big restriction for a Hydroponic system but are sufficiently well understood. Until recently measuring the presence of salts in water electrically was the only way to measure the presence of nutrients in water for commercial growers. Pure water does not conduct electricity. The presence of dissolved salt compounds as free flowing ions is what creates electrical conductivity and so allows electricity to pass through. This means the lower the level of electrical resistivity the greater the presence of dissolved salts.

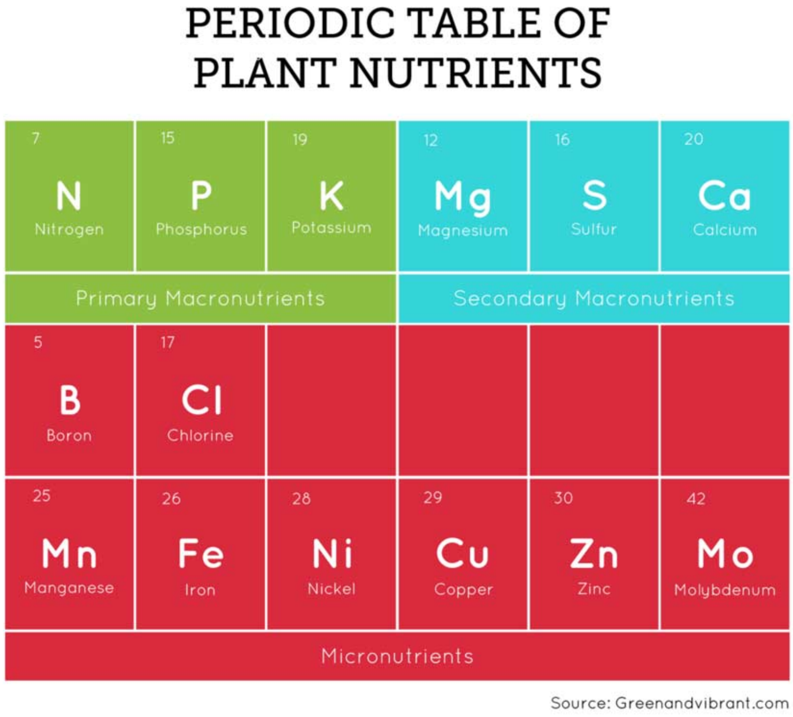
In practice, whilst this offers a simple means of measurement, it is a very blunt instrument and assumes the ratio of differing salts is correct and remains constant. In practice this is not what happens. In modern glasshouses the growing of vine crops is often carried out in gutter systems where nutrient rich water is delivered direct to each plant with roots anchoring the plant in inert media such as mica or rockwool. This should be frowned on as mixing vegetation and mineral wools create waste products that cannot be reused and are very difficult to recycle. Excess water is allowed to drain along the gutters and until recently was simply allowed to drain away and was one of several contributors to nutrient runoff into water courses. Nowadays many regions have legislation that requires the water to be collected and so is reused if possible. Using EC as a measure cannot identify the uptake of individual elements and so the water builds up unwanted salts that can hide a lack of or alter a plant's ability to absorb required nutrients. This means that most who use this method flush their systems every two or three weeks.

Using EC as a measure of total salt levels makes this very difficult and more of an art than a science. There has been much research about the EC level for individual crop varieties. Experienced capsicum growers (often referred to as sweet or bell peppers) use levels measured in milliSiemens of 2-2.5 or 1500ppm. The return water can be as high as 3.5 or 2500ppm. This shows the plants are either consuming more water than nutrients or that there is significant evaporation. What it doesn't show is the usage of individual nutrients.

The level of individual nutrients a plant needs vary according to its stage of growth. Most commercial suppliers of liquid based plant foods sell two main solutions. One is formulated for strong leaf growth and the second promotes flowering and fruit development. As in most natural things plants are far more complicated than this with several other factors impacting on healthy growth.

It should be remembered that, much like animals, the primary aim is to keep the species going. Reproduction is only one part of this. Some plants like cacti have little need to reproduce and many other plants can enter a period of dormancy to survive a winter. Temperature and length/quality of daylight are often the trigger for changing a plant's life cycle stage. Flowering and fruiting are often triggered by identifying the day length reducing. Species that cannot survive a winter (or summer drought) and need to develop seeds that can survive a period of dormancy have all the necessary DNA elements to recreate a new plant.

It has been found that plants have three levels of nutrient dependency for healthy growth [28]. Figure 12 shows three levels of nutrients needed but only shows the nutrients that are taken up via roots alone and makes no mention of carbon, hydrogen and water. Carbon is an essential element for plant growth and mostly collected from air as CO_2. Obviously the other two elements make up water and can be taken up by roots or alongside carbon in the form of air born moisture and CO_2.



PERIODIC TABLE OF PLANT NUTRIENTS

7 N Nitrogen	15 P Phosphorus	19 K Potassium	12 Mg Magnesium	16 S Sulfur	20 Ca Calcium
Primary Macronutrients			Secondary Macronutrients		
5 B Boron	17 Cl Chlorine				
25 Mn Manganese	26 Fe Iron	28 Ni Nickel	29 Cu Copper	30 Zn Zinc	42 Mo Molybdenum
Micronutrients					

Source: Greenandvibrant.com

Figure 12. Plant Nutrient Requirements [28]

When using synthetically produced fertilisers most farmers talk of adding the primary nutrients; nitrogen, phosphorous and potassium (or NPK to use their chemical symbols). For a nutrient to be essential there must a consequence should it not be present. Figure 13 shows how this can be apparent in leaves. This is only a quick snapshot and the lack of lesser needed nutrients can be identified in similar ways. What is most interesting for CEA is that these visible signs of deficiency can be identified with cameras long before the human eye. This is a great example of how CEA can be used to inform traditional agriculture. Plenty Inc. with their huge number of cameras and inconsistent lighting would find this difficult. Done properly with crops passing a smaller number of cameras frequently means the more expensive cameras used in research can be utilised to a far greater extent. Obviously identifying an issue is only the start but in a good CEA facility there is a considerable amount that can be learned about when and how to correct a deficiency or even if it would prove worthwhile for field grown crops. Using collected data with machine learning and artificial intelligence can maximise the benefits of this.

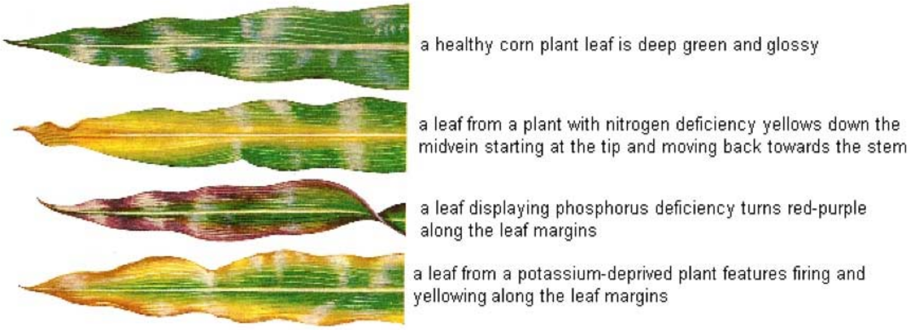

a healthy corn plant leaf is deep green and glossy
a leaf from a plant with nitrogen deficiency yellows down the midvein starting at the tip and moving back towards the stem
a leaf displaying phosphorus deficiency turns red-purple along the leaf margins
a leaf from a potassium-deprived plant features firing and yellowing along the leaf margins

Figure 13. Nutrient Deficiency in leaves. [29]

Greater emphasis is now being shown to the third level of nutrients. There is also the need to combine this with human nutritional needs and so the time has come to move on from using EC as a proxy. Despite this there are still some research programs depending on EC. It should be argued at the very least that this should only ever happen as a means of comparison to help commercial growers who cannot yet justify the investment in new systems.

4.2 Human Dietary Needs

With the current vogue for millennials and others to question food production methods and push for a vegetarian or even a vegan diet there is no better time to identify what we humans needs for a healthy diet. There are some chemical elements that are essential for the human diet that are not needed for a healthy plant. There are many references that suggest there are anything between eight and 14 elements required in small amounts to maintain a healthy body. Not all of them quote the same elements yet offer reasons for each element. One of the most comprehensive listings can be found on the Food and Agriculture Organisation of the United Nations website [30].

The list of elements here is much larger than seen elsewhere. Some of the key points mentioned include:

- Much of Europe is short of dietary selenium which impacts on the ability to make use of other elements.
- Deficiencies of copper, molybdenum and chromium have serious impacts and yet have not been measured until recently.
- The necessity for several elements including cadmium, lead and mercury are of primary concern with the dangers of overexposure and yet have a significant presence in agricultural land in many parts of the world.

One key part should be quoted directly:

New trace elements. *Boron, chromium, manganese, nickel, tin, vanadium, molybdenum, arsenic, lithium, aluminium, strontium, caesium and silicon are regarded as new trace elements in the sense that they have only recently been considered essential in human diets. These elements are the subject of exciting research in animals, particularly ruminants, where they have been shown to be essential in one or more species. For example, ruminants feeding on grass grown in soil where molybdenum levels are abnormally high have demonstrated an increased tendency to exhibit copper deficiency. However, for many of these new trace elements, such as manganese, there is no evidence at the present time that abnormally low or high dietary intakes cause substantial nutritional problems in human populations.*

Thus it can be seen that as agriculture has become increasingly intensive with synthetic growth accelerants (fertilisers) added in large quantities the lesser needed elements have mostly been ignored. The ability to measure the uptake of trace elements has only recently been possible to measure at a price that a farmer would contemplate. The need to do this has never been greater. The removal of lead from road fuels was linked to a substantial drop in minor crime. As the recognition of mental health issues and increasing dependence on prescription drugs rises up political agendas it is likely that links will be found to poor and ill-informed diets [31].

Nutrient	Macro/micro	Uptake form	Mobility in Plant	Mobility in Soil
Carbon	Macro	CO_2, H_2CO_3		
Hydrogen	Macro	H^+, OH^-, H_2O		
Oxygen	Macro	O_2		
Nitrogen	Macro	NO_3^-, NH_4^+	Mobile	Mobile as NO_3^-, immobile as NH_4^+
Phosphorus	Macro	HPO_4^{2-}, $H_2PO_4^-$	Somewhat mobile	Immobile
Potassium	Macro	K^+	Very mobile	Somewhat mobile
Calcium	Macro	Ca^{2+}	Immobile	Somewhat mobile
Magnesium	Macro	Mg^{2+}	Somewhat mobile	Immobile
Sulfur	Macro	SO_4^-	Mobile	Mobile
Boron	Micro	H_3BO_3, BO_3^-	Immobile	Very mobile
Copper	Micro	Cu^{2+}	Immobile	Immobile
Iron	Micro	Fe^{2+}, Fe^{3+}	Immobile	Immobile
Manganese	Micro	Mn^{2+}	Immobile	Mobile
Zinc	Micro	Zn^{2+}	Immobile	Immobile
Molybdenum	Micro	MoO_4^-	Immobile	Somewhat mobile
Chlorine	Micro	Cl^-	Mobile	Mobile
Cobalt	Micro	Co^{2+}	Immobile	Somewhat mobile
Nickel	Micro	Ni^{2+}	Mobile	Somewhat mobile

Table 1. Plant Nutrients and their mobility [25]

4.3 Nutrient Measurement Techniques.

Having noted the comments about nutrients in Hydroponic systems it can be seen that using electrical conductivity should not be considered adequate for growing crops for a healthy human diet. Figure 14 shows more about how nutrient deficiencies can be identified visually. These are pointers we humans can see. As the price of cameras drop we can use dedicated cameras with particular filters to see some of these deficiencies long before they are visible to the human eye. Modern imaging techniques used with contemporary cameras can automate this.

Spectroscopy could also be used. This can identify chemical elements in water and so could be used non-invasively and does not require sacrificial sensing tips. EC sensors have a finite life and need regular cleaning. Measuring pH levels is even more complex with sensor tips rarely lasting more than 6 months with any accuracy.

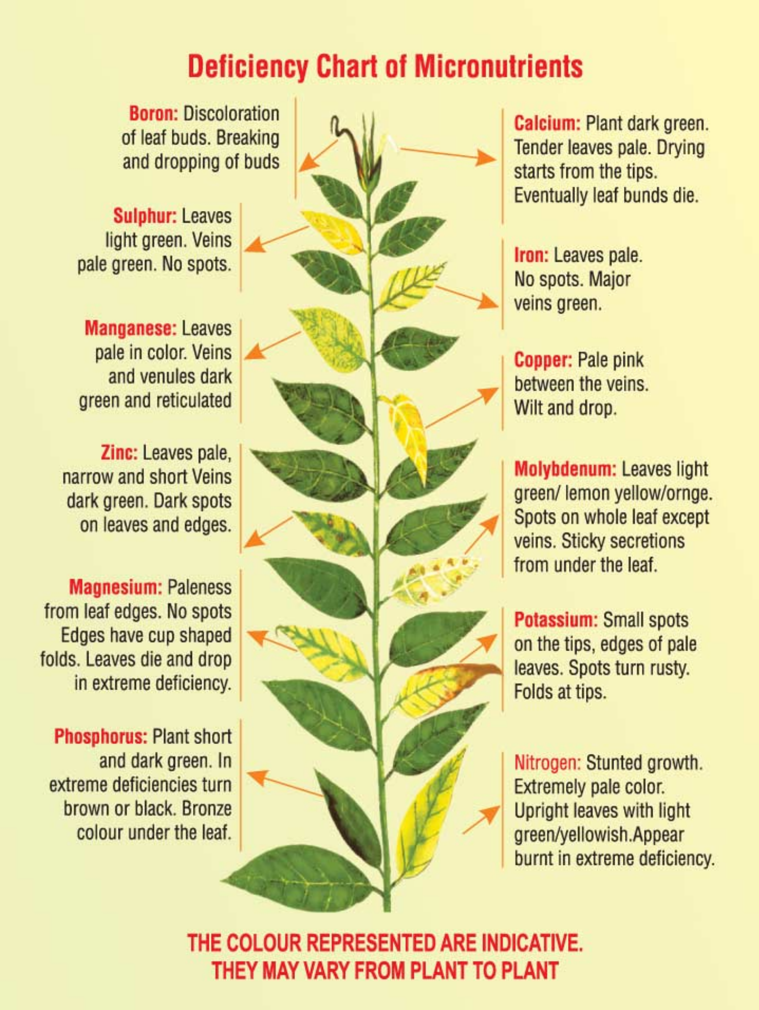

Deficiency Chart of Micronutrients
Boron: Discoloration of leaf buds. Breaking and dropping of buds
Calcium: Plant dark green. Tender leaves pale. Drying starts from the tips. Eventually leaf bunds die.
Sulphur: Leaves light green. Veins pale green. No spots.
Iron: Leaves pale. No spots. Major veins green.
Manganese: Leaves pale in color. Veins and venules dark green and reticulated
Copper: Pale pink between the veins. Wilt and drop.
Zinc: Leaves pale, narrow and short Veins dark green. Dark spots on leaves and edges.
Molybdenum: Leaves light green/ lemon yellow/ornge. Spots on whole leaf except veins. Sticky secretions from under the leaf.
Magnesium: Paleness from leaf edges. No spots Edges have cup shaped folds. Leaves die and drop in extreme deficiency.
Potassium: Small spots on the tips, edges of pale leaves. Spots turn rusty. Folds at tips.
Phosphorus: Plant short and dark green. In extreme deficiencies turn brown or black. Bronze colour under the leaf.
Nitrogen: Stunted growth. Extremely pale color. Upright leaves with light green/yellowish.Appear burnt in extreme deficiency.
THE COLOUR REPRESENTED ARE INDICATIVE. THEY MAY VARY FROM PLANT TO PLANT

Figure 14. Micronutrient deficiencies [28]

Measuring nutrient levels has to become an essential part of growing food for a healthy nutritionally balanced diet and once this data is collected AI can take this learning to a new level. As the desire for more ethically sourced food increases with many turning away from meat towards a plant based diet the need to understand how we can ensure we source much needed trace elements becomes increasingly important. Table 1 shows a listing of what is considered necessary for healthy crop development.

As is so often the case it is not enough to simply look at individual elements as nutrients are taken up as compounds. Again, like the human body some are essential for fixed framework (bones) and others as conduits for nutrients (blood).

4.4 Nutrient Levels in Healthy soils

This chapter has several references to the idea of using AI to learn from CEA and to use that to inform traditional agriculture. Several points have been used to show that CEA should not be thought of as a new or different sector. Hydroponics is a relatively recent word but can be seen to be no more than what plants have always depended on. CEA should be able to do far more to create valuable data that can inform traditional agriculture and so this section has been included to show both the importance of this and what, if done properly, AI can take the data collected at very low cost in CEA and apply the learning for the benefit of all farmers. This can be especially useful in understanding what makes soils healthy which is a subject rapidly rising up the agenda for agriculture as a whole. When using synthetic fertilisers most farmers talk of adding the primary nutrients; nitrogen, phosphorous and potassium (or NPK to use their chemical symbols). The need to look at soil health has largely been caused by issues such as nutrient runoff and compaction seen as key indicators of poor health. Nutrient runoff has undoubtedly been caused by the steady compaction of soil and the related reduction in ability to retain rainwater. Academics such as Professor Simon Blackmore [33] suggest soil compaction is caused by the size and weight of agricultural machinery. This is certainly part of the problem but it can be seen shown that healthy soils recover from the harm of heavy machinery rapidly.

In the UK’s Thames Valley there is a traditional fair that travels the area from late spring onwards. At one particular location near Maidenhead they aim to power the whole event with all electrical power being generated by traditional Showmen’s steam engines. This weekend event takes place in early May and can burn through 8 tons of coal as 6 or 7 Showman’s engines power the stalls and rides. It is said these engines, some well over 100 years old, are working as hard as they would if driving up a steep hill for the whole day. They rock backwards and forwards as they generate the power whilst sitting on little more than a few small boards to prevent them completely sinking in to the soil of an area of protected grass meadow that is noticeably soft and spongy to walk on. Within a month or two of the event there is little sign the land has endured such a tough weekend. So why is this meadow capable of holding so much rainwater and doesn’t suffer from compaction? A heavy steam engine working hard in the same spot for a weekend is far harsher than an arable field having a modern tractor pass over it a few times a year. The difference has to be the health of the soil. This land does not get ploughed every year. This has proven to be a key reason for the drop of biological matter as birds eat the worms brought to the surface and exposed soil carbon oxidises. So what this meadow does not suffer from is the continual industrial scale removal of biological matter which contains all the nutrients a crop needs in addition to the three primary nutrients found in chemical fertilisers. Another reason for soil compaction has to be the demand of intensive agriculture to work the land regardless of soil condition.

As the pressure to increase productivity increased the focus moved from quality to quantity. Now that sensors have dropped in price the time has come to measure a far greater range of trace elements and learn what can be done to help rebuild soil health. AI will play a significant part in learning how this can be done.

North Wyke Farm, the Devonshire farm operated by Rothamsted Research [34], focusses on livestock. One part of their work looks at improving grasslands for grazing livestock. In some of their fields they are able to record data for several parameters and are steadily building a big data set that can be used to predict the best time for remedial treatments. By recording rainfall and the amount of water that flows from a drained field the data can record how, when and in what quantity nutrients flow through the soil. For those registered it is possible to access this data [35]. It is incredible that a rural farm can collect more data than most CEA facilities who make claims about data but do little more than record the time when lights or fans are switched on or off.

5. What can AI offer CEA – and Beyond?

Having looked at many aspects of CEA it can be seen just how complex it is and how much each part influences another. Attempting to control one aspect can be seen to have unintended consequences in other areas.

As in so many areas of life the only way to understand a process starts with measuring. The author has likened factory farming to Heathrow airport. It is well known that this is one of, if not the, most highly stressed airports in the world usually working at over 98% of theoretical maximum capacity. Most factory farms are setting themselves up to depend on the same highly stressed levels and, just like Heathrow, they are working at the extremes of theoretical capacity yet do not know how much they can sell each day. When building an expensive farming facility why would you not take note of dramatic seasonal demand changes? This is not just a problem for areas like the UK where winter demand for salad crops halve. It also applies to places like Florida where summer temperatures are so high many abandon the area for somewhere cooler. Whilst the issue is caused by opposite seasons the need to address the issue remains. Many warm climes rely on polytunnels or more expensive facilities not simply to control the temperature but to protect delicate crops from wild weather events such as hailstorms or dust laden winds. Once the decision has been made to grow undercover the connection to traditional agricultural land changes. There is no better time to change the business model which is what most CEA operators miss. As so often data should be the driver to understand this. Taking a cost saving approach does not preclude the use of contemporary technology but does suggest a more intelligent approach to using it.

If unnecessary costs are avoided in CEA facilities, then so the stress levels can be reduced. Going back to the USDA chart (Figure 1) for romaine lettuce it can be seen that each retail outlet needs a far smaller quantity than the trucks that are bulk carrying them to each state. If we look at modern IT systems and how they use Software as a Service (SaaS) and distributed systems, then why not distribute the growing of fresh produce and offer the monitoring and care management via the cloud? It was this thinking that led to City Farm Systems trademarking CloudGro® for their systems and CloudGrown® for the produce grown in them.

All heavily occupied buildings pay to dump heat and CO_2 – mostly at roof level. They also need a constant supply of food. Rather than use expensive buildings as farms we should look at an alternative method. In India they use rooftop greenhouses as a form of air conditioning. In areas of high heat and poor ambient air quality the cost of filtering and cooling air to make it safe for internal use can be very expensive. By exhausting 'stale air' into a rooftop greenhouse they are able to use plants to consume the CO_2 and release oxygen to refresh the air. This reduces the amount of particles in external air that need filtering out and only needs the temperature to be corrected. At times the need to cool the air is also reduced. Now consider what would happen if the plants were not just ornamentals but a food crop that could be consumed within the building.

The quantities of salad and herb varieties consumed in a building are relatively small daily amounts. Factory farms talk of being able to grow all year round. So why not avoid both the complex and expensive supply chains and the expensive city based factory farms and grow small daily quantities in a rooftop greenhouse? We know that factory farms are set up to grow at capacity but that consumers have varying levels of demand. This is where data can really play a part matching supply and demand. We know that growing in protected environments can grow more than needed locally. We know that current supply chains are complex and wasteful. By growing in smaller quantities at the point of need the time of harvest can be buffered to match demand. Rather than paying to homogenise a whole growing environment it is cheaper to take advantage of zones with differing conditions. We know growth can be slowed in many ways. Reducing CO_2, temperature, water or nutrients are all ways to achieve this and, when programmed to best match a point in the plant life cycle, can be used to improve desired qualities. So as long as the crop planning aims to start with the ability to grow more than expected demand crops can be given alternative growing conditions that alter the rate of growth to match short term demand variations and improve desired qualities.

All of these possibilities need AI to optimise crop management and to match predicted demand back to short cycle growth cycle start points. Rather than follow the example set by arable farmers we need to look further than maximising supply and aim to analyse demand side data, match supply and demand and reduce the costs and efforts needed to achieve this. The opportunity to reduce wastage and, therefore, cost is enormous. At the same time this depends on being able to have customer and harvest within minutes of each other.

Fortunately, the use of AI, sensors, and cameras means this can all be automated and monitored remotely. Rather than having expertise at every location AI and machine learning offer the ability to oversee, manage and offer horticultural expertise via the cloud. With cameras and sensors able to see issues long before the human eye this process offers more than an experienced horticulturist could achieve by walking amongst crops on a daily basis. This is where the possibilities get really interesting.

Whilst most think of CEA as a new horticultural sector City Farm Systems prefers to think it should be a logical extension of growing delicate crops under cover. Throughout this chapter there are examples of how CEA can be improved to be more sustainable and economically viable. Growing at the point of need in low cost facilities is the only way this can be achieved in an expensive city. By doing this properly there is an enormous amount that can be learned to inform and improve many aspects of agriculture as a whole. Automating a greenhouse and using the latest sensing technologies offers the ability to do much more. Using low cost mechanical components to move crop trays around a greenhouse and back to a treatment and sensing area offers far greater potential than the overly complex approach of the high cost operators. Bringing moving crop trays to the point of watering offers the chance to use differing nutrient mixes for each tray. Taking this a stage further with the latest sensing techniques offers the chance to measure nutrient uptake on a regular basis without needing to use invasive testing.

6. Conclusion: The Potential to use CEA derived AI in other areas of Agriculture

The topic of big data has been around for a while. Rothamsted Research are creating big data at their North Wyke Farm [31]. However, there is very little being done to collect crop specific big data in indoor farms where it should be much easier. Rather than creating big data sets, and learning what can be achieved, the overwhelming majority of CEA facilities are paying to create conditions they believe are best for a particular part of the growth cycle for a narrow range of crops. They are then locked into current business models with many stress points.

Using free sunlight wherever possible in automated greenhouses that can be installed in locations that have little alternative use offers the ability to create big data whilst growing food at much reduced cost. Big data is all about gathering data from a broad spectrum and then mining the data to better understand what is happening or to learn something new. Taking a very narrow approach means the outliers are missed. Yet that is where the value is often found. We know plants have developed to cope with most of what nature throws at them. Trying to grow in narrowly defined spectrum of conditions is restricting some of desired qualities. Some are now saying they can cause or trigger specific qualities by changing light recipes. Nature tells us this is not sufficient. Those that research plant specific circadian rhythms show the correlation to light but not that light is necessarily the cause. The ability of plants grown naturally to predict and prepare for dawn goes against the thinking of some artificially lit facilities who want to grow through the night or vary daylength according to energy prices. Even if this made sense there is still the need to vary light recipes mimic the spectral changes of light during the day. The suggestion that crop breeders should breed specific cultivars to match the light qualities a new grower thinks they should offer is not something the author agrees with. The last few decades of breeding have mostly been about increased yield and more recently about reducing dependence on chemical interventions.

The really interesting part is where an alternative approach can lead. We know there is much to be learnt about returning intensively farmed soils to health and improving a crop's ability to match human dietary needs. This is now a point of focus for research. However agricultural research has long depended on manually intensive and slow processes. Much crop breeding and other agricultural research is carried out in greenhouse facilities that our grandparents would recognise. Automated data collection and crop care enabled by an automated greenhouse can remove a huge amount of manual effort and potential error. This means a PhD research student could achieve much more. There would be no need for daily attendance and enable the researcher to have weekends and annual leave – (or time off when unwell). The amount of data would be greater and the risk of inaccurate or missing data due to manual collection would not be an issue. This also means the number or variety of growing experiments can be increased.

Figure 15. Manual tagging and collection of data can lead to errors.

There is also the opportunity to carry out better research with a reduced number of trials. Researchers could be given access to data collected from commercial installations. Mining the bid data would offer the ability to learn from a far wider data set. The result would be learning when and how to take advantage of a plant's growth characteristics.

There is also potential to use this ability in other areas of research such as suggested in chapter 4 where Martínez-España et al. talk of using AI to predict the uptake of carbamazepine and diclofenac in reclaimed water-irrigated lettuces. Prediction is a very important tool but has to be based on understanding what has been found in reality. With an automated system soil samples from polluted or exhausted fields could be used and data collected to improve the modelling. Better still crop varieties could be grown alongside each other. The same variety could be grown with differing treatments or differing varieties could be grown in identical conditions to learn about the differences. It is likely that some varieties will react better than others or that the same variety may be able to cope with poor quality water at one point in its growth cycle but not at another – or at what point this would have a negative impact on harvestable quality. Understanding this could reduce the cost of improving water quality or how to mitigate against naturally occurring variations.

When it comes to breeding crops for CEA growers we also need to think about going back to varieties that were discarded for field based horticulture. Disease is at worst much less prevalent and, ideally wholly excluded, from controlled environments. Mildew is easily treated with UVC light provided human workers are not close by. This suggests cultivars that are prone to these issues could have a place in CEA rather than try to breed variations from existing disease resistant. There is the need for more input from nutritionists and dieticians and a return to growing crops for qualities other than size, colour and to survive supply chains.

Sadly these benefits are something that too many CEA practioners have made impossible with their costly approach to growing food. Just as in many industries the application of AI can help deliver better results.

References

[1] Dr Dickson Despommier, *The Vertical Farm: Feeding the World in the 21st Century,* St. Martin's Press, New York, 2010

[2] Josef Krupicka, Petr Sarec, Petr Novak, *Measurement of Electrical Conductivity of Fertilizer NPK 20-8-8,* Czech University of Life Sciences Prague, 2016 www.tf.llu.lv/conference/proceedings2016/Papers/N237.pdf

[3] TiMet, Horizon 2020 funded research project, 26th September 2014, https://ec.europa.eu/programmes/horizon2020/en/news/study-reveals-plant-growth-ticks-circadian-rhythm

[4] William B. Smith Jr, started the Lean Six Sigma approach when at Motorola, 1980 https://en.wikipedia.org/wiki/Six_Sigma

[5] Professor Marian Rizov, now at Lincoln University, UK http://www.businessdictionary.com/definition/minimum-efficient-scale.html

[6] Article in The Guardian newspaper 21st October 2013 https://www.theguardian.com/business/2013/oct/21/food-waste-tesco-reveals-most-bagged-salad-and-half-its-bread-is-thrown-out

[7] Article from Mail Online 21st October 2103 https://www.theguardian.com/business/2013/oct/21/food-waste-tesco-reveals-most-bagged-salad-and-half-its-bread-is-thrown-out

[8] UK Charity WRAP, http://www.wrap.org.uk/

[9] Rachel Plawecki, Rich Pirog, Adam Montri, and Michael W. Hamm, Comparative Carbon Footprint Assessment of Winter Lettuce Production in two Climatic Zones for Midwestern Market, https://www.canr.msu.edu/resources/comparative_carbon_footprint_assessment_of_winter_lettuce_production

[10] Grist.org: https://grist.org/food/no-lettuce-is-not-worse-for-the-climate-than-bacon/

[11] Eric L Adams, Brooklyn Borough President, *Opening Keynote speech,* Indoor Agtech Conference, New York, June 1028 https://indooragtechnyc.com/venue/

[12] Catherine Luu, *Pharmacy to Farm program*, Nutrition Incentive Program Coordinator at New York City Department of Health and Mental Hygiene, 2016

[13] New York 'Health Bucks' https://www1.nyc.gov/site/doh/health/health-topics/health-bucks.page

[14] Manuka honey and fraud 2nd March 2016: https://www.naturalnewsblogs.com/manuka-honey-fraud-jars-sold-around-world-produced-new-zealand/

[15] USFDA, https://www.fda.gov/Food/RecallsOutbreaksEmergencies/Outbreaks/ucm626330.htm, latest update January 2019

[16] Leila Abboud, *Farm labs that grow crops indoors race to transform future of food*, Financial Times, UK 2nd March 2019, https://www.ft.com/content/6a940bf6-35d4-11e9-bd3a-8b2a211d90d5?hubRefSrc=email&utm_source=lfemail&utm_medium=email&utm_campaign=lfnotification#lf-content=241736989:826782018

[17] Professor Wanlin Gao, Beijing, China, Executive Chairman, *International Conference on Smart Agricultural Innovative Development*, 2017 www.cau.edu.cn

[18] K.P.Ferentinos, I.K.Kookos, K.G.Arvanitis and N.Sigrimis, *From Production to the User: Quality Issues for Agricultural Product Chains, Food Traceability, Tagging Systems,* Chapter 8.2 of the CIGR Handbook of Agricultural Engineering- Vol. 6: Information Technology.

[19] Plenty Inc. South San Francisco, USA https://www.bloomberg.com/news/features/2017-09-06/this-high-tech-vertical-farm-promises-whole-foods-quality-at-walmart-prices

[20] Sean Lennon, Head of Tractor Line, New Holland Agriculture, *IAgrE Conference*, held at Rothamsted Research, October 2017

[21] AeroFarms LLC, Newark, USA https://aerofarms.com/technology/

[22] AquaGrove, Oford MI, USA, http://www.aquagrove.com/aquaponics.php

[23] Graphic from Institute for Systems Biology, High School Intern team, 2015, https://baliga.systemsbiology.net/see-interns/hs2015/projects-2/aquaponics/the-needed-revolution-aquaponics-and-its-presence-in-the-classroom/

[24] Fluence by Osram, https://fluence.science/do-plants-use-green-light/

[25] Professor Abdul Mateen Khattak, *PhD thesis: Spectral Filtering for the Regulation of Plant Growth*, Reading University, UK 1999

[26] XL Horticulture, UK, Manufacturers and suppliers of Polytunnel plastic films. https://www.xlhorticulture.co.uk/

[27] Professor Qichang Yang, *Exploration of Technologies for Vertical Farms and Urban Agriculture,* talk at Nottingham University, UK 2014

[28] Greenandvibrant.com, https://www.greenandvibrant.com/cal-mag-for-plants

[29] Cornell University, Brooklyn, New York, USA https://nrcca.cals.cornell.edu/

[30] Food and Agriculture Organisation of the United Nation, http://www.fao.org/home/en/
[31] FAO, UN, http://www.fao.org/docrep/U5900t/u5900t05.htm
[32] Image Courtesy of: University of Arizona Cooperative Extension
[33] Simon Blackmore, Was *Director of the Centre for Precision Agriculture*, Harper Adams University, UK (now at Earth Rover, an agricultural robotics company)
[34] Rothamsted Research: Orr R, Murray P, Eyles C, Blackwell M, Cardenas L, Collins A, Dungait J, Goulding K, Griffith B, Gurr S, Harris P, Hawkins J, Misselbrook T, Rawlings C, Shepherd A, Sint H, T. Takahashi, Tozer K, Wu L, Lee M (2016)
[35] The North Wyke Farm Platform: effect of temperate grassland farming systems on soil moisture contents, surface run-off and associated water quality dynamics. European Journal of Soil Science 67: 374-385: http://resources.rothamsted.ac.uk/farmplatform

Agriculture and Environment Perspectives in Intelligent Systems
A. Muñoz and J. Park (Eds.)
IOS Press, 2019

doi:10.3233/AISE190003

Intelligent Environmental Biomonitoring Systems: A Promising Arena for Future

Lakshmi GOPAKUMAR[a,1], Ammini JOSEPH[b]
[a]*Research scholar, School of Environmental Studies, Cochin University of Science and Technology, Kerala, India-682022*
[b]*Professor (Retired), School of Environmental Studies, Cochin University of Science and Technology, Kerala, India-682022*

Abstract. Intelligent environmental systems are emerging as one of the most efficient technologies to monitor the effects of contaminants and environmental changes on various species of organisms. As these systems are computer supported, the retrieval of information is very easy and time saving compared to conventional methods involved in biomonitoring. This chapter gives an overall idea of environmental biomonitoring with the inclusion of new advancements in this field. In this chapter the intelligent environmental systems and their importance is given focus with a critical analysis of this new technology in environmental biomonitoring research and analysis. The evolution of intelligent environmental systems from basic lab analysis to incorporation of smart technologies in biomonitoring are discussed. The history of development of intelligent environmental systems shows that environmental biomonitoring is now a highly sophisticated field and has the potential to provide a large volume of information with the aid of modeling and Geographical Information Systems (GIS). As biomonitoring technologies become more and more sophisticated, more studies can be done in the future on plants, animals and human beings without wastage of time and human energy. The chapter mainly targets researchers, students and the people who are interested to know about the new trends in environmental research and development.

Keywords. biomonitoring, computer, environment, environmental information system, environmental management, smart systems

1. Introduction to biomonitoring

Biological monitoring (biomonitoring) is the regular application of biological assessment techniques to get the information regarding the quality and condition of a biological system [1]. Biomonitoring is a variant of environmental monitoring is which organisms are used to monitor the environmental conditions and sometimes predict environmental perturbations based on the data obtained from the study organisms. While environmental monitoring relies on data obtained from physicochemical measurements (for example air quality assessment and water quality determination), biomonitoring relies on data collected from the response of living organisms to a certain condition prevailing in the environment in which they are surviving. Therefore, the concept of bioindicator is an indispensable part of biomonitoring. According to Markert et al. [2], a bioindicator is an organism, part of an organism or community of organisms that contain information regarding the quality of environment. Hence monitoring species are the organisms which help to determine the impact of pollutants or any other external factors causing disturbance to the survival and health of that specific species (test organism). As biomonitoring can perform as a good method to study the environmental variations, in environmental research studies biomonitoring is a means to obtain complete picture regarding the effect of stressors on an organism or group of organisms.

Biomonitoring research studies have developed as a tool to study the health of different ecosystems using bioindicators as agents. Usually biomonitoring studies involve studying a whole organism (earthworm, microarthropods), part of an organism (cell, tissue or organ), or chemical changes produced in an organism in response to any changes in external environment conditions. Its application is highly significant in monitoring the effects of various contaminants on an organism, which is beyond the scope of chemical analysis. Hence the application of biomonitoring is an important part of environmental research and further developments and specializations in this field will enable the development of studies relating to effect of various changes on living beings on the environment. This will help in prediction of future environmental changes and adopt mitigation measures.

The chapter describes the evolution of environmental information systems and its various levels of development into intelligent environmental monitoring systems. The different sections start with application of biomonitoring in environmental studies, in which the concept of biomonitoring is described in detail. This section is followed by current approaches in biomonitoring describing the common laboratory tests in biomonitoring using model organisms. The next sections explain about environmental information systems and its role in biomonitoring followed by the use of biosensors in biomonitoring. The use of intelligent environmental systems and its development in relation with biomonitoring studies is described in the next section. Section 6 deals with intelligent environmental systems, their evolution and current developments. Section 7 and 8 mention the drawbacks and future research perspectives in biomonitoring using intelligent environmental systems.

2. Application of biomonitoring in Environmental Studies

Biomonitoring is a good tool to assess the state of environment, state of a certain organism or community, predict the future state of environment, organism or communities thereby helping in the environment management decision making process. The information regarding the state of organism will gradually point to the health of different ecosystems like air, water and soil. Thus biomonitoring offers a good scope of getting the overall idea related to environment and its associated phenomena. This includes seasonal variations, changes in soil, water and air. Long term biomonitoring also gives information regarding ecological process like food webs and biogeochemical cycling process in the environment. In addition to the study of environmental phenomena, biomonitoring is also a good tool to study the physiological and biochemical response of an organism to a single stressor or an array of pollutants. For example, the effect of various heavy metals, pesticides and insecticides can be studied on a short and long-term basis. Biomonitoring also provides information regarding an organism by analysing the changes occurring in the body of an organism. Hence biomonitoring has widespread application in environmental assessment, life cycle study, physiological, morphological and anatomical studies related to different organisms. In addition to these biomonitoring is an inseparable part of bioremediation research using different organisms especially microbes like bacteria. Biomonitoring also finds application in agriculture and climate change related studies. Therefore, biomonitoring can be considered as an inseparable part of environmental research.

The method of biomonitoring involves the use of different organisms as species for biomonitoring. The information regarding different organisms used in biomonitoring can be seen in table 1.

Table 1. The details of organisms used in biomonitoring

Microorganisms	Plants	Animals
Bacteria	Lichens	Butterflies
Fungi	Bryophytes	Soil microarthropods
Algae	Herbs	Earthworms
Flagellates	Shrubs	Polychaetes
Protozoa	Gymnosperms	Clams
		Fishes

Among the various organisms mentioned in Table 1, bacterial flora has been used as excellent tools of biomonitoring. Among plant group, lichens have been considered as one of the traditional plant groups which can monitor air pollution especially sulphur dioxide pollution in the atmosphere. Polychaetes have been used for marine pollution monitoring and earthworms have been considered as model organisms to monitor the effects of soil pollution. A few species of earthworms like *Eisenia foetida* and *Eisenia andrei* have been used to study the effects of various pollutants like organic compounds, pesticides and heavy metals on soils. A few toxicity tests are usually carried out using these organisms as a part of biomonitoring studies. The most popular tests include acute toxicity tests and chronic toxicity assays. In all these assays the end points are different and are species specific, the common end points of acute toxicity tests being mortality and end points of chronic toxicity tests being reduction in biomass and reduction in reproductive capability.

3. Current approaches in biomonitoring, Pros and cons of traditional methods

The concept of using of different organisms for estimating environmental stress was a new concept during the early stages of biomonitoring research. At the beginning of biomonitoring research, biomonitoring was widely used in streams because streams were considered to be the most endangered systems in the world. The earliest tool for biomonitoring in aquatic ecosystems uses the concept of saprobian system which indicates the biologically decomposable organic pollution in running water ecosystems [3]. The saprobian system relies on the saprobic values of various insect species between a scale of 0 and 4. Even though saprobian system relies on niche concept and considers abundance and numbers of various species, it cannot be applied across large geographical areas and for ecosystems having various ecological impacts [3]. During its early stage of development, the concept of biomonitoring was widely applied to streams using periphytons and fishes [4]. Among the periphytons, diatoms have been most extensively used in which chlorophyll a and biomass have been used as the measures of environmental stress. Later, benthic invertebrates have also been used as indicators for biomonitoring water ecosystems. Compared to fishes and periphytons, benthic invertebrates have the advantage of representing site-specific environmental conditions due to their sedentary nature. They also act as a connecting link between the abiotic and biotic components in the water ecosystems. But a high number of biomonitoring studies have been conducted later using algae. The advantages of using algae for biomonitoring was their high sensitivity to pollutants, ease of sampling and cosmopolitan distribution with popular autecology [5]. During the later period, the incorporation of various fields in biomonitoring has resulted in the development of biomonitoring as a separate branch in environmental research. The current trends in biomonitoring research involves uses a combined approach of laboratory methods and field sampling to estimate the degree of response in a living organism to certain pollutant or stress. For example, soil biomonitoring uses bacteria, fungi, microarthropods, annelids and nematodes to study the effects of contaminants and stress on living organisms [4]. Similarly, water analysis uses fishes and algae for estimating stress in water ecosystems. The concept of biomonitoring which has started with river biomonitoring using various organisms has gradually developed into human biomonitoring in which the human tissue samples like blood, urine and milk are studied for the presence of various chemicals [6]. Currently, these are mostly used for occupational health related studies. Generally, the earlier biomonitoring research concept relied on the number and species composition of different organisms as an indication to the changing environmental condition of a certain area. An example is the increased number of algae in a pond due to eutrophication and reduced number of earthworms in soil due to chemical pollution. But in many cases such a relationship may be ambiguous and insufficient to give a clear picture of a scenario. A new concept has been put forward by European Water Framework Directive (WFD) called Danish Stream Plant Index (DSPI) [7]. The various aquatic plant traits studied using DSPI were ecological preference, life forms, morphology and dispersal of each plant species. DSPI is considered as a good tool to study the anthropogenic impacts of pollution and management on aquatic ecosystems [7]. The recent trends in biomonitoring research involves the use of bioassays and biomarkers [8, 9]. Both of these concepts are almost similar. While a bioassay is an experiment conducted to determine the toxic effects of a pollutant/ pollutants to a certain species of organism, a biomarker is a compound in the body of an organism which has the ability to fluctuate in response to various environmental conditions. Bioassays include acute toxicity assays, chronic toxicity assays and reproduction assays. The most common biomarkers studies include lipids, neurotransmitters and various proteins in an organism like stress proteins. The biomonitoring studies in fishes especially lake trouts considers lipids, age, length, weight, sex, growth, diet and bioenergetics for estimating the effect of polychlorinated biphenyls (PCBs). The recent techniques involve use of different models and spatial assessment for data analysis [10]. For a detailed study of the effects of various stressors, biochemical analysis of the organisms and their body parts becomes a necessity. This involves killing of the organisms for laboratory studies. The drawbacks of the current approach involve extensive sampling, taxonomic identification tedious laboratory assays and mortality of the test organism. Most of these toxicity assays also involve an optimum period of more than 10 days to obtain the results. For example, The Organisation for Economic Co-operation and Development (OECD) recommends acute toxicity tests for a minimum period of 14 days using earthworms to get the results [11]. OECD also suggests more than 60 days for chronic toxicity tests [12] using earthworm assay· The recent approaches in river biomonitoring involves the use of functional measures like nutrient uptake, secondary production and molecular techniques as more precise and specific tools for biomonitoring [4]. Another recent advancement is the use of traits in biomonitoring. A trait is a characteristic which gives information regarding the species adaption to an environment and are divided into ecological traits and biological traits [5]. The main aim of trait-based biomonitoring research is establishing relationships between traits and environmental factors. In addition to these, multimetric and multivariate approaches have been used in which statistics has been incorporated as a strong stool for biomonitoring research [3]. Feeding guilds and multiple biological traits are also included which makes the concept of biomonitoring more efficient compared to the conventional methods. A new concept which was introduced in biomonitoring studies are studies using exposomes. Exposome is an environment complement to the genome determining the risk of disease [13]. An exposome can give information regarding chemical and non-chemical exposures which an individual accumulates from its pre-natal development [14]. As exposomic research can include all exposures of potential health significance, they can perform a number of untargeted analysis which can measure a number of low abundant metabolites present in an organism, which is impossible in conventional biomonitoring research. The detection of unknown chemicals using exposomic research can be further strengthened with the help of bioinformatics [13]. Seasonal observation and sampling of the test organisms is a major issue makes conventional biomonitoring a tedious task. In order to solve these issues in scientific world, intelligent systems are a good option. The following chapter describes how intelligent systems can be effectively used in environmental biomonitoring.

It is seen that traditional biomonitoring relies on the use of various organisms and indices and establish the relationship between population dynamics and environmental conditions. This has certain advantages. The main advantage is that it gives the information related to present field conditions in an ecosystem. Another advantage is that the techniques are mostly observatory, excluding the use of tedious mathematical calculations and predictions. A major disadvantage is the lack of universal applicability of a technique. For example, the saprobian index has limited geographical applicability as a major disadvantage. There are a number of disadvantages to the primitive methods in biomonitoring. One among them is that the population and species distribution of a region may be dependent on more than a single factor which may not be included in the list of environmental parameters to be assessed. Also the dose-response relationship between various parameters cannot be properly and precisely established using the traditional methods like saprobian index. To rectify this, the use of functional group concept and biological group concept are included in present biomonitoring methods. Additionally, the incorporation of various biomarkers and bioassays are helpful in quantifying the toxic levels and tolerable levels of each toxicant or groups of toxicants in an organism. The introduction of multivariate statistical methods and models at present make biomonitoring more scientific and gives reliable results compared to conventional methods.

4. Environmental Information Systems (EIS)

Environmental information systems can be considered as a bridge between biomonitoring and environment intelligent systems. Environmental information systems describe a class of systems which performs environmental monitoring, data storage, disaster response, reporting, simulation, modeling and decision making [15]. The need for accuracy and ready availability of data regarding environmental responses and organisms has demanded the incorporation of information technology in environmental management systems. This can be achieved by the inclusion of ‘informatics’ in environmental information systems. The concepts like bioinformatics [16, 17] and ecoinformatics [18, 19, 20] has already emerged as ’smart methods’ to monitor environmental data compared to conventional methods. But all these are in a nascent stage of development and requires further specialization. As environmental biomonitoring is a broad area which includes proper documentation and dissemination of environmental data using living organisms, the incorporation of informatics in environmental biomonitoring information system will facilitate provision of real time and accurate data over a period of time. The incorporation of informatics can be done both at laboratory and field level. At the laboratory level this is achieved through automated continuous monitoring systems and specifically at field level it can be done through the incorporation of Geographical Information Systems (GIS) in data collection and analysis. Both these approaches are relevant depending upon the type of monitoring studies. For example, microcosm and mesocosm experiments using fish, plants and invertebrates use continuous monitoring systems while the collection of field data involves the use of geographical information systems. Depending upon the purpose and scope of biomonitoring, GIS can be incorporated at various stages of experimentation and documentation. The overview of GIS components is represented in Figure 1. It is seen that the use of environmental information systems can be helpful for disseminating more data compared to conventional methods of biomonitoring which relies only on laboratory analysis and field observations undertaken by human beings.

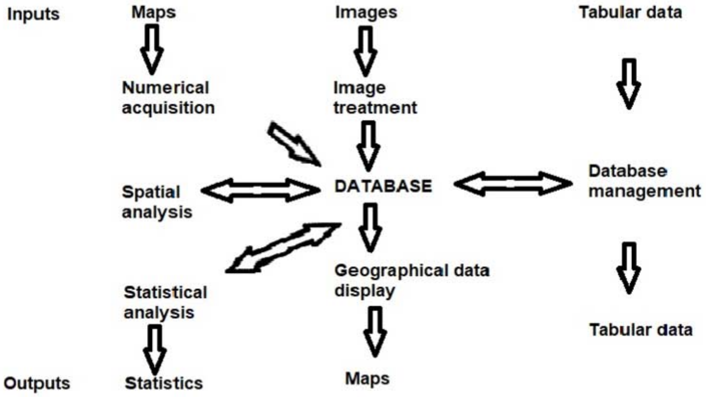


Figure 1. Overview of GIS components

Figure 1 explains the integration of various GIS components used in environmental applications. The various inputs like maps, images and tabular data are given into a computer. These together with numerical acquisition, image treatment and spatial analysis forms the database of environmental biomonitoring. After data processing using statistical analysis and geographical data display, the outputs are given as maps and tabular data of geographical information systems. This same principle is followed during biomonitoring using Geographical Information Systems.

A further advanced stage of research in environmental information systems is the application of intelligent systems. The biosensor concept has emerged to be much important in studying the response of certain organisms to environmental fluctuations. A biosensor normally uses the sensitivity of biological components like biomolecule or cell coupled with transducers to provide information regarding the state of an organism by measuring certain parameters of the cell or molecule. A biosensor works on the principle that any change in the biological and physiological response in an organism can be converted into electronic signal through a transducing system, which can quantify the response of the organism. Normally physiological and biochemical activities are measured using biosensors. The advantages of biosensors are low cost, hassle free use, continuous monitoring ability and high specificity [21, 22]. These are widely used for studying the effects of industrial emissions and xenobiotics on an organism. In a biosensor mechanism, the binding of an analyte to the target molecule results in a change of physicochemical properties like pH, electron transfer etc. Hence genotoxicity, immunotoxicity and endocrine response can be studied through a biosensor system [23, 24, 25]. Biosensor concept is widely applied in healthcare research, especially in the manufacture of handy self-use instruments like glucometers. In environmental research field, biosensors are mainly used in microbiological research in which bacterial cells are incorporated in the biosensor system. An example of a sophisticated biosensor is Automated Water Analyser Computer Supported System (AWACSS) which can simultaneously measure certain organic pollutants [26, 27]. The implementation of biosensor research in environmental research field [28] has given the idea regarding the use of real time systems in biomonitoring and finally to intelligent environmental systems.

5. Intelligent environmental biomonitoring systems

Intelligent environmental systems are the next step of development after biosensor technology which can be addressed as an initiative based on ‘smart technology’. Intelligent systems are the monitoring systems which facilitate recording of responses in an organism with the aid of computational, mathematical and statistical analysis. Sometimes these are also used to predict the future responses or fate of a certain condition on a test organism.

Environmental information systems generally use a sequence of several components which is a combination of physical, chemical and biological elements. An environment intelligent system usually consists of a biological chamber, a recording mechanism which can continuously record the movement of an organism, and an instrument which can give the output of observation. The advantage of the use of artificial intelligent systems over primitive systems are ease of collection of experimental data and recording of results. The schematic representation of an intelligent environmental system is given in Figure 2.

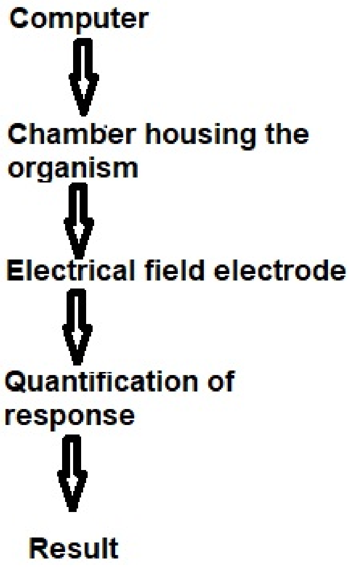

Computer
Chamber housing the organism
Electrical field electrode
Quantification of response
Result

Figure 2. Schematic representation of an intelligent environmental system

Usually the intelligent environmental system consists an electronic device like microchip, a chamber to keep the test organism and a computer which can continuously monitor the organism's response towards a certain condition and record the data based on the response. As real time data can be obtained without much hassle and with the click of a button, these systems can be considered as handy while compared to the traditional biomonitoring systems. The intelligent environmental monitoring systems are 'really smart' systems as they are less time consuming, more accurate in computation and applicable in real time as well as simulated environments in environmental biomonitoring studies. The main difference of these systems over the conventional methods is reliable data collection regarding various parameters without including much human labour.

Intelligent environmental systems are a multidimensional approach involving incorporation of physical, chemical, biological and computational analysis in environmental information systems which can give real time information related to a certain environmental condition. The use of intelligent environmental systems in biomonitoring facilitates the evaluation of current status of an environment/ ecosystem in relation to a living organism/ part of organism**.** In other words, it is a good option to get an overall picture of environmental variations and an organism's response in relation to certain conditions. The Figure 3 shows the relevance of environmental biomonitoring systems as a combination of various fields acting as an efficient tool for research and day to day applications.

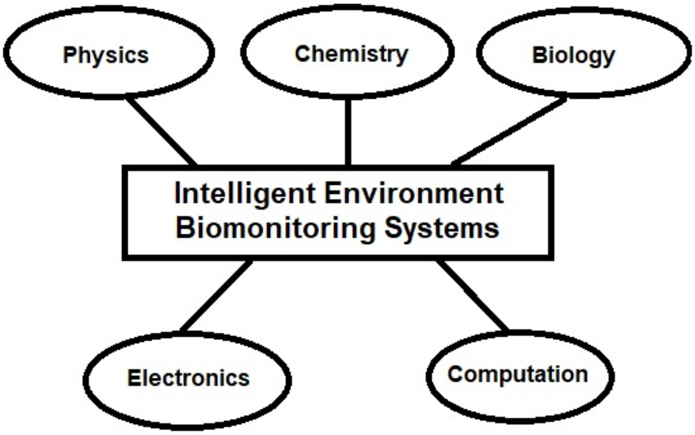


Figure 3. Multidimensional approach in intelligent environmental biomonitoring

As the figure depicts, intelligent environmental monitoring systems uses a combination of various specialized fields to get the complete information regarding a certain situation. This is advantageous as it eliminates major difficulties involved in conventional research approaches and gives a comprehensive pool of data which gives large volume of information.

There are a number of advantages when we compare conventional biomonitoring methods and intelligent environmental biomonitoring systems. As intelligent systems are computer controlled, the most important benefits are accuracy, precision and large volume of data when compared to conventional biomonitoring approaches which includes extensive sampling together with lab and field analysis. Table 2 gives the comparison of conventional biomonitoring methods with intelligent environmental biomonitoring systems.

Table 2. Comparison between conventional biomonitoring approaches with intelligent system approach

Conventional biomonitoring	**Intelligent system biomonitoring**
Time consuming(days to months)	Very less time consuming(seconds to minutes)
Not very accurate	Accurate
Human errors can affect the original results	Human errors have little or no influence on the original results
Continuous monitoring may not be possible	Continuous monitoring is possible even at seconds' level
Laboratory based	Can be incorporated into real time environment
Involvement of humans is mandatory for experimentation	Human involvement is not mandatory as automated systems are incorporated for deriving results
Relies on biological and chemical analysis	Relies on biological, chemical, computational analysis and robotics

The comparison among conventional approach with intelligent environmental systems shows that the latter is much effective, less time consuming, more precise with a large volume of data output.

6. Intelligent environmental systems- evolution and current developments

The evolution of intelligent environmental systems was a rapid development within a short period of time with the support of information technology. The branch of intelligent environmental systems has strong basement on environment information systems. Environment information systems, together with laboratory analysis and field observation approach developed into a network of information gathering sources, which facilitated easy data recording. When information technology developed introducing the concepts of virtual space and cloud computing together with the advent of nanotechnology, intelligent environmental systems developed further as an innovative and recognized field in information technology- supported scientific studies. The development of intelligent environmental systems is illustrated in Figure 4.

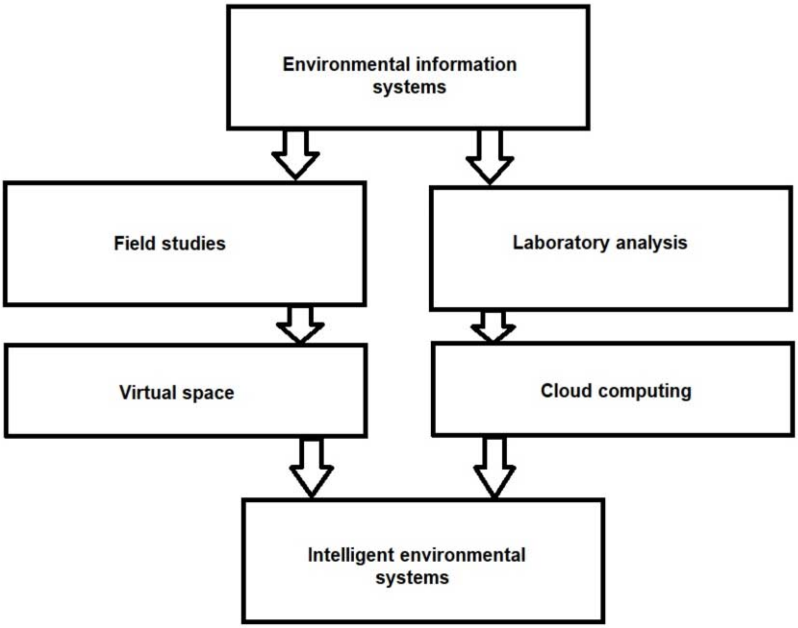

Environmental information systems
Field studies
Laboratory analysis
Virtual space
Cloud computing
Intelligent environmental systems

Figure 4. Evolution of environmental systems [29]

The development of information technology has resulted in two important concepts: Internet-of-things (IoT) [30] and cloud computing [31]. Both of these became advantageous to intelligent environmental biomonitoring systems. The internet of things has evolved into a system using multiple technologies by 2013. The Internet-of-Things

[32, 33] includes a network of physical objects connected with embedded electronics, sensors, software and networking which enables to collect data. This in fact, refers to a network which can collect data and which can be controlled and programmed. The technologies range from micro- electromechanical systems to embedded systems. This gives the opportunity to integrate physical objects to computer based systems. The advantages of these systems are increased accuracy and efficiency [34]. Cloud computing is an internet based computing which provides data to computers and other devices. The concept of cloud computing arose in 1950, as the companies required multiple user access to certain information from various regions. The concept has evolved from multiple concepts like grid computing, utility computing etc. but the real concept of cloud computing materialized only as early as 2007. Cloud computing [35] enables computation in a virtual space with the help of internet. The advantages of cloud computing in intelligent environmental systems are low cost, ease of access, increased storage space and ease of access to information regarding biomonitoring parameters. The incorporation of modeling in computer based systems and the use of algorithms will further give reliable results which is applicable for prediction of future environmental conditions with the aid of various software. The algorithms will give step wise instructions regarding the actual processing of the data fed into the systems based on user needs and modeling helps to predict the results of a certain environmental observation (for example carbon emissions) for the coming years based on the present scenario. The introduction of Decision Supporting Systems (DSS) is a good approach for modeling studies related to biomonitoring. The concept emerged during mid-1960s when computerized models were used for decision making and planning. A DSS is a computer-based problem solving system which uses the concept of modeling to support decision making process and is of two types. An active DSS is an integrated DSS which suggests possible actions and inform about the criterion used while a passive DSS provides only information. A more advance version of DSS is IDSS (Intelligent/Integrated Decision Supporting Systems) [29]. An IDSS can also answer logical questions through simulation models. An example of IDSS is AIDAIR [29] which was developed in Geneva for air quality management. AIDAIR uses GIS core and used for urban pollution management. It has three different components like TAP (Traffic and Air pollution); Energy GIS (EGIS) and Air Pollution and Public Health (APPH). All these are interconnected. The AIDAIR system was integrated with information layer which contains the information related to air quality bioindication with lichens [29]. In this method, the air pollution was monitored with the Index of Air Pollution (IAP) calculated from number of certain types of lichens over tree trunks, modelling and visualizing the results using a GIS based approach. With this method, the IAP of an area for a certain period of time can be monitored and visualized, thus facilitating the continuous monitoring of atmospheric pollutants, and thereby the air quality of a given area.

As intelligent environmental system is a newly emerged concept, only a few studies have been carried out in the field of environmental biomonitoring utilizing intelligent system technology. This is because this field is relatively young compared to other scientific fields and is currently in a developing stage. Most of these studies have used a certain organism/ species of organism for monitoring various parameters of air and water. After reviewing the previous works in this area, a summary is shown in Table 3.

Table 3. Biomonitoring research works involving intelligent systems

Name of the system	Organism monitored	Reference
AIDAIR decision support systems	Lichens	Degli-Agosti et al., 1998[29]
ECOTOX	Flagellates	Tahedl and Hader , 1999[36]
Effluent biomonitoring system	Fish	Shedd et al., 2001[37]
Multispecies freshwater biomonitor	Amphipoda	Kirkpatrick et al ., 2006[38]
Automatic biomonitoring system	Insect	Houng and Chou., 2012[39]
Lab-on-a-chip platform for rapid toxicity test	Artemia	Huang et al., 2016[40]
Biochip -G and biosensor module	Algae	Umar et al., 2017[41]

The different 'smart' biomonitoring systems as obtained from the literature are discussed below.

When we consider intelligent environmental systems for biomonitoring, the most popular one is the automated fish biomonitoring system for determining effluent toxicity which was a successful venture [37]. Similarly, a multi-agent system called O3RTAA was designed for monitoring air-quality. The system utilizes data coming from a meteorological station. In this system, several software agents co-operate to monitor both meteorological and air pollutants. The data obtained from several sensors were collected for assessing air-quality. The system used alarms to give warning to recipients based on the parameters studied [42]. This system can be represented as three layers involving three agents. The upper distribution layer with distribution agents for alarm distribution to users. The second layer is management layer with alarm agents for data analysis, storing and alarm identification. The third layer is the contribution layer with diagnosis agents for measurement validation and estimation of missing values. Information flows from field sensors through these three agent layers. The diagnosis agents capture the air quality measurements received into the system and deliver them to the management layer after validation process. The alarm agent receives these measurements and decides whether an alarm should be given. The validated measurements are stored in the system future use. A water monitoring system was introduced as automated water quality monitoring system to study the response of animals in relation to water quality parameters [39]. The system consisted of a chamber housing an aquatic organism. Ventilatory behavior and body movement were recorded to identify the effect of stressors on the movement behavior of the organism. The responses were recorded by a sensor which gave the result regarding the stress of the organism [42]. Further developments of monitoring systems include fully automated video data analysis to study the toxic effects of water pollutants and behavioral changes of the aquatic organism. Thus the technique becomes more user friendly, with the visualization of real time data with the aid of video camera [40].

The development of Biological early warning systems has emerged as a method to quantify and monitor responses in animals. This has been developed using fishes as experimental organisms for biomonitoring studies [43, 44]. The opercular movement and gill purge in fish ventilatory systems have been used to study the response of fishes to stressors. An example is the fish biomonitoring in which fishes exposed to effluent was monitored. A continuous automated biomonitoring system was set up utilizing the ventilatory movement of the bluegill *Lepomis machrochirus* at Abradeen Proving Ground [37]. In this system, water from a contaminated aquifer will be pumped into the treatment facility to remove heavy metals and organic pollutants. Subsequently, the treated effluent will be pumped into one of the two holding tanks and a side stream will be directed to a biomonitoring chamber with fish. The holding tanks has a mechanism to stop effluent discharge in case of fish mortality. Temperature, pH, dissolved oxygen and conductivity of water will be measured in the control and test chambers every 30 minutes and ventilator movement of fishes will be measured using electrodes suspended above and below each fish. The electrical signals will be amplified and passed to a computer for analysis. A multispecies freshwater biomonitor has also been developed which used impedance conversion to study vertebrate behavioral responses [30]. This study tests the possibility of use of crustacean *Corophium volutator* for sediment biomonitoring. The multispecies freshwater biomonitor is based on quadropole impedance conversion. The system had two chambers (pair chamber) joined together and connected to multispecies freshwater biomonitor and personal computer. The lower chamber had bioban-spiked sediment and upper one had overlaying water. The sediment was prepared in two concentrations and the paired chambers were placed in individual cylindrical transparent plexiglass chambers. The experimental organism *C. volutator* was placed within each pair of chambers and left for 20 minutes to settle and recordings were made over a period of 1 hour. Behavioural pattern of the crustacean could thus be monitored and recorded.

An important feature of the smart biomonitoring systems mentioned above is that most of them are based on water biomonitoring and air quality monitoring. There is very little development of intelligent systems related to soil pollution research [46, 47, 48]. These were mostly used for the remediation of petroleum contaminated soils, estimation of spatial distribution of heavy metals in the soil and as early warning systems of risk assessments of heavy metals. The methods for continuous monitoring of effects of soil pollution on soil fauna is still at a developmental stage but offers immense opportunities in the future. The development of ‘smart’ intelligent systems mostly used biosensor technology to record the response of organisms. Also most studies were recent as this field of research is still in a young stage of development.

7. Drawbacks of intelligent environmental biomonitoring systems

The use of intelligent environmental biomonitoring systems is considered to be advantageous due to its benefits but there are certain drawbacks for the usage and implementation of intelligent environmental biomonitoring systems. The most important one is the problem in implementation. A good intelligent biomonitoring system requires sufficient lab space exclusively used for experimental set up. For the biomonitoring studies using higher organisms, more space is required. The requirement of good infrastructure is an essential component for implementing intelligent environmental systems in biomonitoring. Hence only good laboratories in developed countries will be successful in implementing intelligent environment biomonitoring systems. The developing countries will not be able to contribute much to this area due to financial constraints. The design of software for biomonitoring has to be species- specific. Therefore, intelligent biomonitoring systems require specific instrument design and software development according to the nature of different species. Hence proper care is needed during the design of software which can control automated intelligent biomonitoring systems. The software has to be perfect, including all measurable parameters, user friendly and free from errors. The second problem is the costs involved in the design of such systems. The costs involved in the purchase of software can be high which will make these systems inaccessible to the researchers and usually only big industries may be able to afford them. As a solution to this problem, software has to be developed using open source platforms. Even though the biomonitoring system may be 'smart' and automated, trained personnel will be needed to monitor the system for detecting any faults in order to guarantee reliable results and ensure continuous working of the system. This requires thorough study of the instrument set up and proper troubleshooting. Another problem is the limited development of these monitoring system at present. Only a few organisms were monitored until this time using intelligent monitoring system. The studies were mostly related to water and air pollution. There are no studies related to soil pollution. Also climate change related studies are not properly addressed.

8. Future research perspectives

The future research in intelligent environmental biomonitoring systems seems promising as it may revolutionize the present method of biomonitoring in scientific and industrial analysis. While the conventional biomonitoring research relied on continuous sampling and laboratory analysis, the future research will be carried out with the aid of information technology, bioinformatics and molecular methods. The future systems will be handy incorporating electronics, computation and biological methods. The major advantage will be real time visualization of data using modelling software. The studies will be focused on various monitoring organisms at species and ecosystem levels. With the incorporation of open source software platforms, small microcosms like aquariums and terrariums can be set up in laboratories with a computer to control and record the behavior of the organisms in response to environmental and climatic changes. The next stage of development will be portable intelligent environmental biomonitoring systems which can be transported into various research stations, labs and research vessels. There is also scope for development of portable air and soil biomonitoring systems using smart technology. This will also help to predict climatic changes in the future leading to revolutionization in climate change research. Also non accessible areas like deep seas and river bottoms can be well studied and the faunal response can be documented. This will be particularly beneficial in the study of various algal species in the marine and freshwater ecosystems. There is also scope for identification of new species in inaccessible areas like bottom seas. A breakthrough in the field of biomonitoring research will be the advanced levels in exposome incorporated biomonitoring where all the stressors of an organism can be identified within a short time span. The molecular methods will also show further advancement by incorporating metagenomic research using various genetic materials in different organisms, from bacteria to human beings. All these will foster and facilitate the study of ecosystems throughout the world. Hence more data can be created on the impact of various changes in the environments thereby strengthening further research studies. These data can be used for environmental protection, ecosystem restoration, environmental prediction and environmental management.

Acknowledgements

The authors acknowledge School of Environmental Studies, Cochin University of Science and Technology, Kerala, India. We also thank University Grants Commission, India for giving financial support for the research work.

References

[1] John C. Environmental biomonitoring, assessment, prediction, and management: certain case studies and related quantitative issues. Food and Agricultural Organisation, 1979.

[2] Markert BA, Breure AM, Zechmeister HG. Bioindicators and biomonitors. Elsevier, United Kingdom, 2003.

[3] Bonada N, Prat N, Resh VH, Statzner B. Developments in aquatic insect biomonitoring: a comparative analysis of recent approaches. Annual Review of Entomology.2006; 51, 495-523. doi:10.1146/annurev.ento.51.110104.151124.

[4] Li L, Zheng B, Liu L. Biomonitoring and Bioindicators Used for River Ecosystems: Definitions , Approaches and Trends. Procedia environmental sciences. 2010; 2, 1510–1524. doi:10.1016/j.proenv.2010.10.164.

[5] Wu N, Dong X, Liu Y, Wang C, Baattrup-pedersen A. Using river microalgae as indicators for freshwater biomonitoring: Review of published research and future directions. Ecological Indicators. 2017; 81, 124–131. doi:10.1016/j.ecolind.2017.05.066.

[6] Angerer J, Aylward LL, Hays SM, Heinzow B,Wilhelm M. Human biomonitoring assessment values: Approaches and data requirements. International Journal of Hygiene and Environmental Health. 2011; 214, 348–360. doi:10.1016/j.ijheh.2011.06.002.

[7] Baattrup-Pedersen A, Gothe E, Riis T, Andersen DK, Larsen SE. A new paradigm for biomonitoring: an example building on the Danish Stream Plant Index. Methods in Ecology and Evolution. 2017; 8(3), 297-307. doi:10.1111/2041-210X.12676.

[8] Regoli F, Giuliani ME. Oxidative pathways of chemical toxicity and oxidative stress biomarkers in marine organisms. Marine Environmental Research. 2014 Feb 1; 93:106-17. doi:10.1016/j.marenvres.2013.07.006.

[9] Viarengo A, Lowe D, Bolognesi C, Fabbri E, Koehler A. The use of biomarkers in biomonitoring: a 2-tier approach assessing the level of pollutant-induced stress syndrome in sentinel organisms. Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology. 2007 Sep 1; 146(3):281-300. doi:10.1016/j.cbpc.2007.04.011.

[10] Gewurtz SB, Backus SM, Bhavsar SP, McGoldrick DJ, de Solla S R, Murphy EW. Contaminant biomonitoring programs in the Great Lakes region: review of approaches and critical factors. Environmental Reviews. 2011; 19(NA), 162-184. doi:10.1139/a11-005.

[11] Spurgeon DJ, Hopkin SP. Extrapolation of the laboratory-based OECD earthworm toxicity test to metal-contaminated field sites. Ecotoxicology. 1995 Jun 1;4(3):190-205. doi:10.1007/BF00116481.

[12] Robidoux PY, Svendsen C, Caumartin J, Hawari J, Ampleman G, Thiboutot S,Weeks JM, Sunahara, GI. Chronic toxicity of energetic compounds in soil determined using the earthworm (Eisenia andrei) reproduction test. Environmental Toxicology and Chemistry. 2000; 19(7):1764-1773. doi:10.1002/etc.5620190709.

[13] Dennis KK, Marder E, Balshaw DM, Cui Y, Lynes MA, Patti GJ, Rappaport SM, Shaughnessy DT, Vrijheid M, Barr DB. Commentary Biomonitoring in the Era of the Exposome.Environmental health perspectives 2017; 125(4), 502–510. doi:10.1289/EHP474.

[14] Cernansky R. A blend of old and new: biomonitoring methods to study the exposome. Environmental health perspectives.2017; 125(4), A74. doi:10.1289/ehp.125-A74.

[15] Athanasiadis IN, Mitkas PA. An agent-based intelligent environmental monitoring system. Management of Environmental Quality: An International Journal. 2004; 15(3): 238-249. doi:10.1108/14777830410531216.

[16] Van Aggelen G, Ankley GT, Baldwin WS, Bearden DW, Benson WH, Chipman JK, Collette TW, Craft JA, Denslow ND, Embry MR, Falciani F. Integrating omic technologies into aquatic ecological risk assessment and environmental monitoring: hurdles, achievements, and future outlook. Environmental health perspectives. 2009 Aug 17; 118(1):1-5. doi:10.1289/ehp.0900985.

[17] Jones MB, Schildhauer MP, Reichman OJ, Bowers S. The new bioinformatics: integrating ecological data from the gene to the biosphere. Annual Review of Ecology, Evolution, and Systematics. 2006 Nov 7; 37.

[18] Hale SS, Hollister JW. Beyond data management: how ecoinformatics can benefit environmental monitoring programs. Environmental monitoring and assessment. 2009 Mar 1; 150(1-4):227. doi:10.1007/s10661-008-0675-x.

[19] Suri A, Iyengar SS, Cho E. Ecoinformatics using wireless sensor networks: An overview. Ecological Informatics. 2006 Nov 1; 1(3):287-93. doi:10.1016/j.ecoinf.2006.02.008.

[20] Michener WK, Jones MB. Ecoinformatics: supporting ecology as a data-intensive science. Trends in ecology & evolution. 2012 Feb 1; 27(2):85-93. doi:10.1016/j.tree.2011.11.016.

[21] Malhotra BD, Singhal R, Chaubey A, Sharma SK, Kumar A. Recent trends in biosensors. Current Applied Physics. 2005; 5(2):92-97. doi:10.1016/j.cap.2004.06.021.

[22] Kress-Rogers E. Biosensors and electronic noses for practical applications. Handbook of Biosensors and Electronic Noses, Medicine, Food, and the Environment, CRC press, United States of America, 1997.

[23] Rogers KR. Biosensors for environmental applications. Biosensors and bioelectronics. 1995; 10(6-7):533-541. doi:10.1016/0956-5663(95)96929-S.

[24] Chiti G, Marrazza G, Mascini M. Electrochemical DNA biosensor for environmental monitoring. Analytica Chimica Acta. 2001; 427(2):155-164. doi:10.1016/S0003-2670(00)00985-5.

[25] Rodriguez-Mozaz S, de Alda MJL, Marco MP,Barcelo D. Biosensors for environmental monitoring: A global perspective. Talanta. 2005; 65(2):291-297. doi:10.1016/j.talanta.2004.07.006.

[26] Tschmelak J, Proll G, Riedt J, Kaiser J, Kraemmer P, Barzaga L, Jackson M. Automated Water Analyser Computer Supported System (AWACSS) Part I: Project objectives, basic technology, immunoassay development, software design and networking. Biosensors and Bioelectronics. 2005; 20(8):1499-1508. doi:10.1016/j.bios.2004.07.032.

[27] Tschmelak J, Proll G, Riedt J, Kaiser J, Kraemmer P, Barzaga L, Jackson M. Automated Water Analyser Computer Supported System (AWACSS): Part II: Intelligent, remote-controlled, cost-effective, on-line, water-monitoring measurement system. Biosensors and Bioelectronics, 2005; 20(8):1509-1519. doi:10.1016/j.bios.2004.07.033.

[28] Batzias F, Siontorou CG. A novel system for environmental monitoring through a cooperative/synergistic scheme between bioindicators and biosensors. Journal of environmental management. 2007 Jan 1; 82(2):221-39. doi:10.1016/j.jenvman.2005.12.023.

[29] Degli Agosti R, Couach O, Fiore-Donno AM, Clerc P, Dubois A, Courvoisier O, Hussy C, Fedra K, Greppin H, Haurie A. Integration of biological indicators (lichens) of the state of the environment in a spatial (GIS) intelligent decision support system (IDSS). The Co-Action between Living Systems and the Planet, University of Geneva, 1998.

[30] Evans D. The internet of things: How the next evolution of the internet is changing everything. CISCO white paper. 1(2011); 1-11.

[31] Mell P, Grance T. The NIST definition of cloud computing. Recommendations of the National Institute of Standards and Technology, U.S. Department of Commerce, 2011.

[32] Atzori L, Iera A, Morabito G. The internet of things: A survey. Computer networks. 2010 Oct 28; 54(15):2787-805. doi:10.1016/j.comnet.2010.05.010.

[33] Gubbi J, Buyya R, Marusic S, Palaniswami M. Internet of Things (IoT): A vision, architectural elements, and future directions. Future generation computer systems. 2013 Sep 1; 29(7):1645-60. doi:10.1016/j.future.2013.01.010.

[34] Popovic T, Latinovic N, Pesic A, Zecevic Z, Krstajic B, Djukanovic S. Architecting an IoT-enabled platform for precision agriculture and ecological monitoring: A case study. Computers and Electronics in Agriculture. 2017; 140:255-265. doi:10.1016/j.compag.2017.06.008.

[35] Armbrust M, Fox A, Griffith R, Joseph AD, Katz R, Konwinski A, Le G, Patterson D, Rabkin A, Zaharia, M. A view of cloud computing. Communications of the ACM. 53(4), 50-58. doi:10.1145/1721654.1721672.

[36] Tahedl H, Hader DP. Fast examination of water quality using the automatic biotest ECOTOX based on the movement behavior of a freshwater flagellate. Water Research. 1999; 33(2):426-432. doi:10.1016/S0043-1354(98)00224-3.

[37] Shedd TR, Van Der Schalie WH, Widder MW, Burton DT, Burrows EP. Long-term operation of an automated fish biomonitoring system for continuous effluent acute toxicity surveillance. Bulletin of environmental contamination and toxicology. 2001 Mar 24; 66(3):392-9. doi:10.1007/s00128-001-0018-x.

[38] Kirkpatrick AJ, Gerhardt A, Dick JT, McKenna M, Berges JA. Use of the multispecies freshwater biomonitor to assess behavioral changes of Corophium volutator (Pallas, 1766) (Crustacea, Amphipoda) in response to toxicant exposure in sediment. Ecotoxicology and environmental safety. 2006; 64(3):298-303. doi:10.1016/j.ecoenv.2005.07.003.

[39] Houng HY, Chou JM, Development of Automatic Bio-Monitoring System for the Life History of Insect. In Applied Mechanics and Materials (Vol. 195, pp. 1078-1082). Trans Tech Publications, Switzerland, 2012.

[40] Huang Y, Persoone G, Nugegoda D, Wlodkowic D. Enabling sub-lethal behavioral ecotoxicity biotests using microfluidic Lab-on-a-Chip technology. Sensors and Actuators B: Chemical. 2016; 226:289-298. doi:10.1016/j.snb.2015.11.128.

[41] Umar L, Setiadi RN, Hamzah Y, Linda TM. An Arduino uno base biosensor for water pollution monitoring using immobilized algae Chlorella. International Journal on Smart Sensing & Intelligent Systems. 2017; 10(4).

[42] Shedd TR, Widder MW, Leach JD, Van Der Schalie WH, Bishoff RC. U.S. Patent No. 6,393,899. Washington, DC: U.S. Patent and Trademark Office, 2002.

[43] Korver RM, Sprague JB. A real-time computerized video tracking system to monitor locomotor behavior. Automated Biomonitoring: Living Sensors as Environmental Monitors. 1988:157-71.

[44] Kramer KJ, Botterweg J. Aquatic biological early warning systems: an overview. Bioindicators and environmental management. 1991 Sep 26:95-126.

[45] Geng L, Chen Z, Chan CW, Huang GH. An intelligent decision support system for management of petroleum-contaminated sites. Expert systems with applications. 2001 Apr 1; 20(3):251-60. doi:10.1016/S0957-4174(00)00063-4.

[46] Wu S, Zhou S, Chen D. A Framework of Intelligent early warning system for pollution risk of heavy metals in soil. In2010 International Conference on Intelligent System Design and Engineering Application 2010 Oct 13 (Vol. 1, pp. 20-23). IEEE.

[47] Hu DW, Bian XM, Wang SY, Fu WG. Study on spatial distribution of farmland soil heavy metals in Nantong City based on BP-ANN modeling. Journal of Safety and Environment/ Anquan Yu Huanjing Xuebao. 2007; 7(2):91-5.

Agriculture and Environment Perspectives in Intelligent Systems
A. Muñoz and J. Park (Eds.)
IOS Press, 2019

doi:10.3233/AISE190004

A Novel Design and Implementation of Data Acquisition and Preprocessing System for Precision Agriculture

Seonghun Lee
Department of Computer Science and Engineering
Chung-Ang University, Seoul, Korea
gnstjdok@cau.ac.kr

Abstract. Increasing agricultural productivity is a global concern as food security is expected at the risk in the near future. Lots of information technology based studies have shown positive effects in analysing and streamlining the agricultural work. However due to frequent data shortage, it is difficult to facilitate precision agriculture. Then an easy deploying, preprocess-integrated data acquisition system may help to solve these problem. This paper presents a system that generates various temporal data streams and refines them automatically. The system has a server and one or more station. The station is dedicated to a plant and collecting a heterogeneous data periodically. The number of station is easily extendable to gather more crop's information. For effective data collection, all sensors have the same sampling period, 1 *min*. This sampling period is based on the daylight and its mathematical foundation is presented too. Data preprocess is conducted with low-pass filter and resampler. A method to find an optimal filter based on root-mean-square-error is proposed and analysed.

Keywords. Data Acquisition, Data Preprocessing, Precision Agriculture

1. Introduction

National food security has been one of the most important concerns for decades. However, farm population and agricultural production relative to Gross Domestic Product (GDP) have been steadily declining and this makes the food supply in the world unstable [1]. Increasing food resource requires more land and workers, which is not feasible due to significant cost. Therefore it led us to the need for research on increasing agricultural productivity.

The most remarkable advances in agriculture over the ten years has been made by adopting Information and Communications Technology (ICT) [2,3]. Digitalized agriculture with sensors and actuators enables farmers to take advantage of ICT such as detection and prediction [4–6]. However collecting data from farmland and plants is very arduous and difficult task, although the data itself is valuable. In other words, data insufficiency is the major problem of applying information technologies. Many studies have shown that precision agriculture needs a valuable and reliable large amount of time-series data [7–9].

Plant phenotyping is a successful field of information technology that can be applied to precision agriculture and may be one of the pillars of improving agricultural productivity. There are lots of studies for attaining spatiotemporal data of plants and analysing for in-depth use with them [10–13].

A lot of researches already covered the beneficial effects of applying sensor technology and Internet of Things (IoT) to enhance the accuracy and reliability [14–18]. Moreover, it is a foregoing trend to adopt computer vision solution into agricultural area. This trend is referred as high-throughput phenotyping that collects spatiotemporal data of plant and its environment including visual information [19].

However, lots of data acquisition or phenotypic systems do not consider how raw data is susceptible to noise [20, 21]. For example, if a shadow occurs temporarily or the wind suddenly blows, the environmental facts may change. This does not have a significant effect on true observations. Or, unstable power can cause digitization errors, which is undesirable. There are, of course, many studies to minimize digitizing errors, but these studies are limited to solving hardware errors. On the other hand, in order to facilitate the analysis process, it is advisable to appropriately reduce the noise of the observations prior to the data analysis stage. However, it is usually handled by a data analyst that is separated from the data acquisition system.

One of the most widely used methods for noise reduction is a spectrum based noise filtering. Typically, environmental data are distributed over low-frequency area. Then the refined raw data would be attained by taking high-frequency area away. However, determining the appropriate threshold is a tricky task. Inappropriate thresholds cause almost no action or result in significant loss of meaningful data. Therefore, setting a proper cut-off frequency is very important.

This paper presents a design and implementation of data acquisition and preprocessing system for precision agriculture. The system has two processing modules. The one is to make various measurements on the environmental facts of the target plant and the other is to refine collected time-series raw data. In the front-end data acquisition phase, hardware equipments, analog-to-digital conversion (ADC) devices, are used. On the other hand, most of the data filtering and processing steps are fully implemented in software. To denoise collected environmental data, the Fast Fourier Transform (FFT) is applied for spectrum analysis. A method to determine proper cut-off frequency for low-pass filter (LPF) is presented too. This method takes pre-given noise ratio for calculating the optimal cut-off frequency.

Several analyses and experiments are held to determine proper sampling rate and to prove the effect of preprocessing. Also, a method to choose the proper sampling rate is provided. According to the daily sunshine duration on two solstices in Seoul, Korea, 1 cycle-per-minute is proved to be enough as the sampling rate in this paper. On the other hand, to check the effect of filtering process, root-mean-square-error (RMSE) based analysis was performed. The result of filtering shows the effect of LPFs according to various cut-off frequencies. Also, the comparison results according to various LPFs and noise ratios are provided to analyse the noise immunity of each LPF.

The designed system is intended for use on small farms of several thousand square meters or more. The number of data collecting stations can be easily increased. This system is designed to collect various heterogeneous data and provide them into a refined time-series data. Proposing system will also help for any farmer or researcher by easing efforts to adopt the data acquisition system. Any analog sensors can be attached easily into this system with the minimal effort of calibration. Also, the proposed automated preprocess allows data analysts to focus solely on data analysis itself.

2. Data Acquisition System

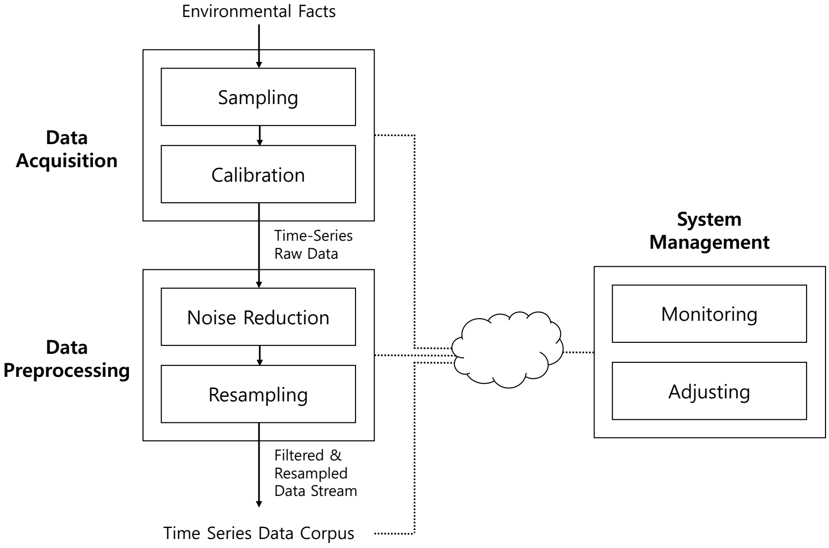


Figure 1. Block diagram of proposed system

The system is designed to collect multiple environmental facts and provide a refined agricultural data stream. All heterogeneous data is collected and integrated as a unified data corpus on this system. The raw data stream is then filtered and re-sampled with prespecified parameter. This system also allows remote monitoring and control of internal system parameters.

The proposed system consists of three subsystems: acquisition, preprocessing and management module (**Figure 1**). Among them, the data acquisition and preprocessing modules are integrated as a data generating procedure that provides a unified, refined data corpus. On the other hand, the management module is operated and processed independently. User can monitor the states of each data generating subsystem and control them in remote with his/her own device.

The data acquisition module generates raw data from its front-end collecting device. The sampling rate is pre-given and can be changed in run-time. All gathered data are also calibrated with their pre-defined calibration tables. The user can easily attach new sensors into this acquisition module if there is the proper calibration information. All sensors in this system uses the same sampling rate for development convenience. However it is possible to apply different sampling rate for each sensor. Also sensors for high-throughput phenotyping purpose can be attached to improve the quality of the dataset. A Charge Coupled Device (CCD) camera or a Light Detection and Ranging (LiDAR) sensor, for example, may provide the shape of a plant, leaf or even the density of plants which can help to figure out more high-level data such as the volume of a leaf or plant [22, 23]. In that case, the sensor should be attached to data acquisition module without using the ADC and the sampling period must be considered properly due to phenotyping cost.

The raw data from the previous module is refined by the data preprocessing module. The degree of refining can be set by the user. Also if the user wants to preserve raw data, the refinement procedure can be skipped. The first step is noise reduction based on LPF. The cut-off band can be determined automatically or manually. Because the characteristics of each of environmental facts are all different, the cut-off frequencies should be set differently for each data stream. After reducing noise, data are sampled again at the re-sampling rate that the user provides. This re-sampling work is supported in the software layer, thus changing the re-sampling rate does not compromise the continuity of data collection, unlike the sampling process of the data acquisition module.

2.1. System Design

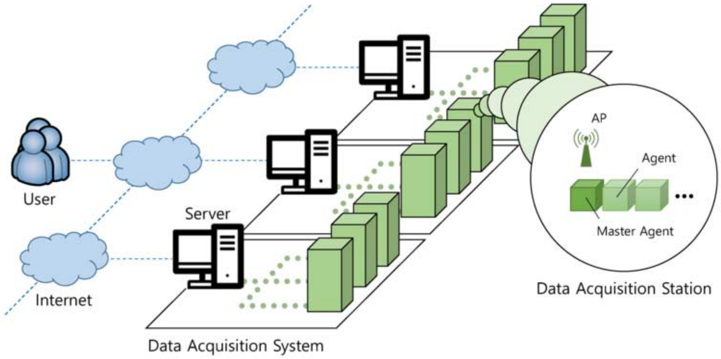


Figure 2. System outline

Figure 2 shows the usage of this system. Proposed system aims to a moderate sized farmland. The system gathers agricultural data that are generated in a farmland and send them to an internal or external server. A system has one or more local networked stations that is dedicated to collect agricultural data streams of a plant. These roles can be nested as needed, such as in small farm. For example, a station can act as a server and configure itself without any other station.

Multiple stations are also effective in a situation where the farmland is large or there are plenty of plants. Of course, it is necessary to install many stations in situation that data of heterogeneous plants should be collected. Different species must be handled in separate stations to obtain their own agricultural data. However it is also effective within a homogeneous plant. This is particularly noticeable when there is a significant difference in system installation area. A farmland, so called as 'macro area', would be uneven so that the environmental facts of the 'micro area' of the system may be different. Further, the gap would be larger in controlled environment such as automatic water dispensed area. Some plants can be overlapped on watering and some are not. This phenomenon causes the non-uniformity of the target data. In such situation, the generalized data corpus can be obtained by averaging the data that are collected from the distributed stations within a system.

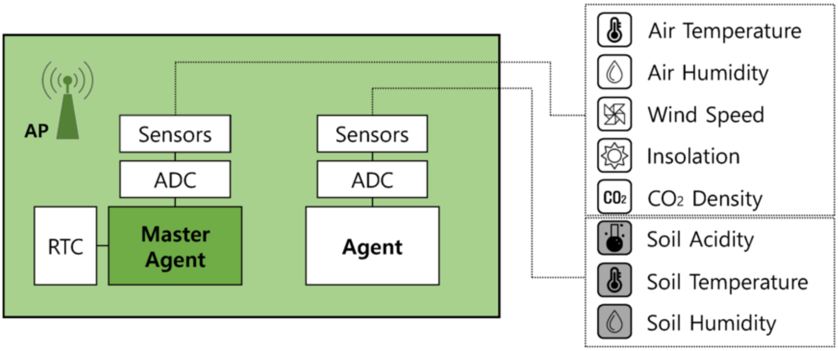


Figure 3. Diagram of the station configuration

Table 1. List of data types and its samples

Data Type	Unit	Sample
Timestamp	YMDHMS	20180701120301
Air Temperature	°C	32.7 °C
Air Humidity	%	38 %
Soil Temperature	°C	21.7 °C
Soil Humidity	%	27 %
Illuminance	lux	9187 lux
CO_2 Density	ppm	412 ppm
Soil Acidity	pH	6.8 pH
Wind Speed	m/s	2.7 m/s

A station can be configured with several agents and sensors. The number of agents may be increased to support sensors or to reduce station workloads. Typically one or two agents are enough to configure feasible data acquisition system. However when system requires more sensors, additional agents could be attached to support the sensors. The agent is OS-supported embedded platform with an ADC.

A station has a master agent that intermediates between a server and a station. The master agent sends a periodic collecting signal to the other agents, retrieves raw data and packages them into a unified data corpus. In addition, the master agent has a real-time clock (RTC) to synchronize all the agents on the network and keep the sampling period accurate and precise.

Figure 3 is an example of a system design. Two types of sensors are connected with the agents: ground and underground. All agents on each station are connected to the wireless AP assigned to each station.

The acquired data corpora are stored to the master agent and sent to the server when the agent is idle. **Table 1** shows a structure of a data corpus. Because the listed types of sensors are adopted to test the scalability of the system, some sensors may be redundant for practical use. On the other hand, these sensors should be properly calibrated with the calibration table provided by the user.

The server collects datasets from stations that are distributed through farmland. As a result, this data accumulation will produce a large-scale data on the underlying environment of the local farmland to enable precision agriculture.

2.2. Sampling

Crop development is affected by various environmental factors. Typically well-trained farmer always considers climate to fertilize and manage crops. But this manual checking is not required to be often because the actions the farmer can take are limited, such as watering and fertilizing. In contrast, precision agriculture requires more than climate information [24]. To optimize and automate the agricultural task, the feature of crops should be extracted from periodically collected environmental data [25].

The problem is how frequently this system collect data. There is a trade-off between the analysis accuracy and the computation burdens including storage size. The best solution is to find the optimal sampling rate where collected data is redundant but also not insufficient. But determining which sampling rate is enough to deploy is another problem.

For example, for the soil pH, one measurement per year may be enough, but for the air temperature, more frequent measurements are necessary. The temperature usually increases from sunrise and decreases after sunset. Thus, the primary frequency is assumed to be 1 cycle per day (*cpd*). To obtain the annual temperature change, one sample per day may barely satisfy the requirement. But it is not enough for practical applications. The dynamic range of the temperature changes in a day is also an important fact when monitoring plant growth. Thus, sampling has to be performed at a much higher rate than 1 *cpd*. By the way, this is the same reason why the daily temperature trend is not visible in the annual temperature trend [26]. If there is the need to attain more precise trends, the number of samples in a cycle should be increased.

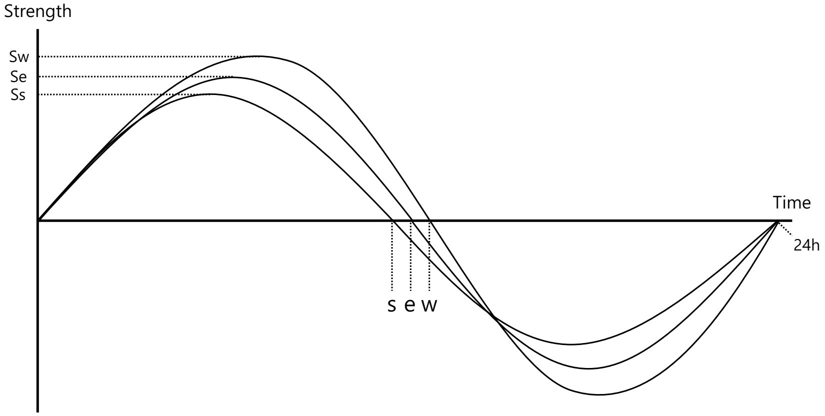


Figure 4. A daylight hours model in a sinusoidal form

For example, 1 *cpd* is considered the primary frequency of the daylight hours. If the insolation is assumed to be zero at sunrise and sunset, the changes in insolation can simply be modeled as a 'half-sine' shape. If the night hours are assumed to be the counterpart of the daylight hours, the daylight hours with insolation can be modeled as a sinusoidal form, as shown in **Figure 4**. When it is the equinox, the shape becomes a sine wave. However, the length of the daylight hours continues to oscillate between the summer and winter solstices. The shape should be modeled not as a pure sine wave but as the sum of the harmonics of sine waves, of which the principle frequency is 1 *cpd*. For accurate daylight hours, therefore, a sampling rate much higher than 2 *cpd* is required.

More specifically, the optimal sampling rate can be calculated by considering different daylights between two solstices. **Figure 4** shows the simplified form of annual daylight shifting. Point 's' and 'w' means each solstice points. Point 'e' represents the equinox. In case of Seoul, Korea, the daylight is long as 14 *h* 47 *min* on the summer solstice. On the other hand, when the winter solstice, it take 9 *h* 45 *min*. There are toughly 5 *h* time differences between the two solstices. Consequently the daylight hour shift per a day becomes $365 \ / \ 10\ h = 1.644\ min$. This shows that if the information that we want to extract is strongly related to daylight, the sampling period should be at least $1440\ cpd \ / \ 1.644\ min = 876\ cpd$. Then 1752 *cpd* will be the Nyquist sampling frequency. However for implementation convenience the adopted sampling rate is 1440 *cpd*, which means collecting on every minutes.

2.3. Data Filtering

Noise is an inevitable problem in a digitization process. This problem comes from the limitation of modern technology and cannot be completely solved. In other word, the digitization phase is naturally noisy. For example, it is very difficult to obtain samples at precisely regular intervals and to quantize very accurately. For example, the *Atmega328P*, which is used in *Arduino*, the tiny computing board, has an 8-bit ADC which has 2 LSB error. So the theoretic minimal error is $2^2/2^8$ = *approx.* 0.38%. In addition, unstable power can make accurate quantization difficult. In the quantization step, the reference voltage is used to quantize the sample, but the unstable and fluctuating power can cause the results to be inaccurate.

However noise does not occur only in the above hardware-intrinsic cases. The environmental facts of the system are also noisy. The weather may change every day. Even on the same rainy day, the time to start raining, the time of rain, and the amount of rainfall are all different. There are also temporary environmental changes such as shading, rapid wind and heat source or obstacle. These changes are temporary and do not make much sense.

Data analysis is highly dependent on data quality. In other words, data corruption has a significant impact on analytical results. As a result, data analysts should reduce noise from the data or use noise-insensitive analysis algorithms. However the latter requires a broad understanding of data analysis techniques. Moreover data analysts should be aware of the characteristics of each data type. Thus providing reduced noise data from data acquisition module eliminates such difficulties. However, this difficulty is eliminated if the data acquisition system provides the noise-reduced data.

Typically, a hardware-based filter is applied in front of the sensor to reduce noise. However, this approach requires prior knowledge of the input signal. It even degrades the portability and scalability of the data acquisition module. In addition, the presence of additional hardware costs more than software, and software developers have difficulty for understanding the entire system mechanism Thus, software-based filtering method is adopted in this paper.

In general, most of the environmental data is distributed in low-frequency bands. Noise, on the other hand, can be in proportion to the original signal or can be span the entire spectrum. This observation shows that a well-chosen LPF reduces lots of noises and minimizes the loss of informative data.

However, another problem is how to set the appropriate cut-off frequency. Inadequate cut-off frequencies may take meaningless computational cost or loss the useful data. If the threshold is too high, most noise will be preserved, so filtering will not work. Conversely, too low threshold can compromise the informative true observations. Therefore, the cut-off frequency should be determined as appropriate.

Generally, the cut-off frequency is determined in manual. When adopting sensors in a system, the user should investigate the signal characteristics and set the proper thresholds with his/her own intuition or formula. In other words, every time a new sensor is attached, or even when there is a significant change in environmental facts, a significant amount of user effort is required.

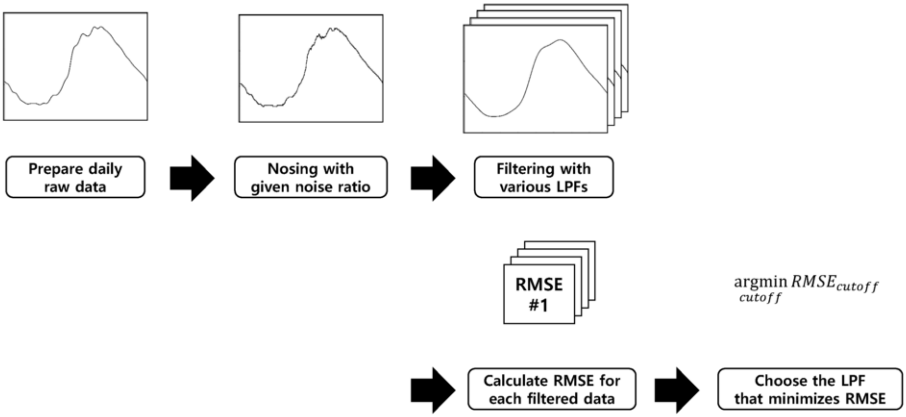

Prepare daily raw data
Nosing with given noise ratio
Filtering with various LPFs
RMSE #1
$\underset{cutoff}{\operatorname{argmin}} RMSE_{cutoff}$
Calculate RMSE for each filtered data
Choose the LPF that minimizes RMSE

Figure 5. Proposed method to set the proper LPF based on RMSE

Therefore, selecting an LPF automatically will help users to handle the system. In this paper, an RMSE-based LPF selection method is proposed as shown in **Figure 5**. This method requires a parameter to represent user's noise ratio assumption. With a given noise ratio, the original raw data is artificially noised and then filtered back over various LPFs. The LPF is determined by the location that minimizes the RMSE between the original data and the filtered data. However, this filtering step is ignored if the prime frequency of the target data is too large for filtering process to make sense.

In the nosing stage, noise is randomly generated as positive or negative amount with given noise ratio in proportion to the original data. Of course, this generated noise is arbitrary and can not be used directly to determine appropriate thresholds. Therefore, it is recommended to use the average of multiple RMSEs obtained from the iterative noise procedure. In the experiment, the minimum number of iterations to get a plausible result is $1,000$.

This method has limitations in that it does not use noise-free data as the original signal. If the informative band is known, the chosen LPF will get more effective. However, since it is extremely difficult to obtain noise-free data, artificial noise is used instead.

After the filtering process, the filtered time-series data may be re-sampled using a pre-given re-sampling rate from the user. In order to maintain the continuity of data stream, the re-sampling phase must be located after the filtering phase. Otherwise, the filtered data may not match the re-sampling period at the moment the user changes the re-sampling rate. In this case, the filtered data will be discarded, which destroys the continuity of the data stream. On the other hand, if a server storage is highly concerned and the importance of data is not high, this re-sampling work can be effective in solving many problems.

2.4. Implementation

In this paper, a station is implemented and tested for research purpose. A total of two agents are employed to test the data acquisition. The *RaspberryPi 3*, a tiny embedded platform, is used as an agent and the *Raspbian Stretch 4.9*, a *Debian*-based OS, runs on each agent. These materials are well-known for their easy operability and usability with powerful resources. It is more expensive than other non-OS-powered board, however it is chosen because OS support makes system development, deployment, and processing easier. The *DS3231* 8-bit RTC chip on the *Pioneer 600 expansion board* was used for time keeping. Also a 24-bit *ADS1256* ADC chip on *high-precision AD/DA board* is selected to sample the sensor value. For monitoring purpose, *RaspberryPi Touch Display* is deployed.

Table 2. List of deployed sensors

Type	Model	Manufacturer
Air Temperature	NTSF-1	Bu Yuan Elec.
Soil Temperature	NTSF-1	Bu Yuan Elec.
Air Humidity	MM2001	Max Detect
Soil Humidity	SEN030003	DFRobot Elec.
CO_2 Density	SEN0159	DFRobot Elec.
Soil Acidity	SEN0161	DFRobot Elec.
Wind Speed	SEN0170	DFRobot Elec.
Photoresistor(Illuminance)	GL20528	Senba Opt& Elec

Figure 6 shows the master unit and the sensor bed that is installed on the experimental purpose frame. As an example of installation, the frame is installed with a square 1 m^3 size. Of course the shape and structure of the frame is changeable to meet the need of user. The hardware frame is equipped with two collecting agents and eight sensors in the experiment. All devices use DC power with no DC voltage regulators. The implementation takes about $200 without sensors. An agent takes toughly $70 except an $30 RTC module. All the sensors are analog to be attached on the ADC. The adopted sensors are listed in **Table 2**. To measure environmental factors, a small pot plant is also chosen. By

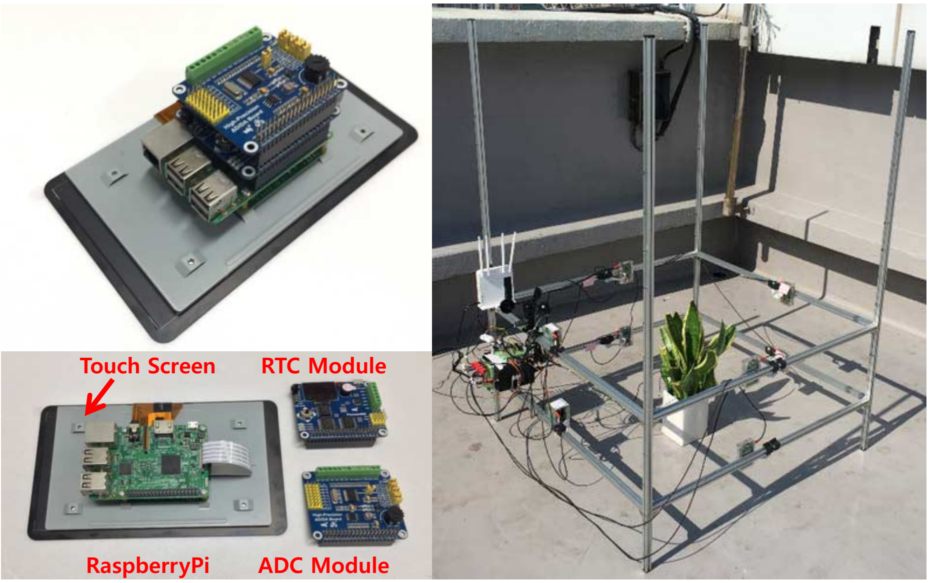


Figure 6. A master agent (Left) and deployed station (Right)

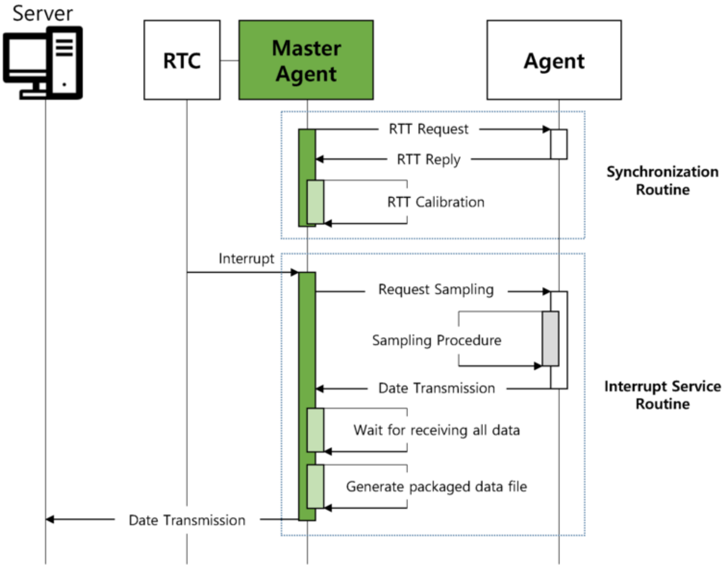


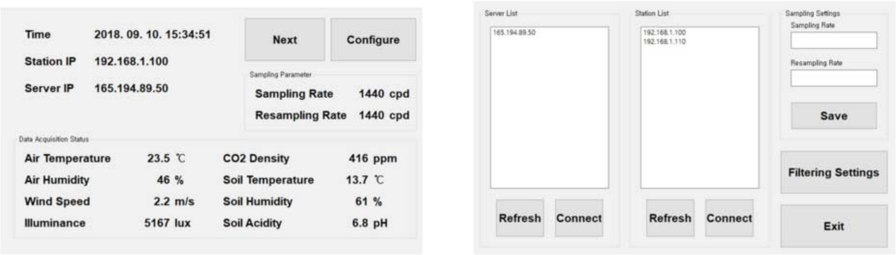
Figure 7. Data collection procedures in the system

the way the equipments in **Figure 6** are more various for performing other experiments, thus it can be ignored.

All participating agents must be well synchronized to collect accurate data. Therefore, the master agent acts as a Network Time Protocol server to evenly synchronize other devices. Moreover, the round-trip-time (RTT) between the master agent and other agents is measured for command arrival calibration. If the RTT is less than the longest RTT of the station, the signal transmission time is compensated by an additional delay of the differential amount.

This procedure is shown as **Figure 7**. Too frequent synchronization may disturb the data acquisition procedure, thus it is manually performed by farmers. After synchronization, every collecting sequences are triggered by the RTC interrupt. Then the master agent acts send sampling request to other agents including master agent itself. Every agents then reply for the request with their own collected data. Finally the master agent packs all the data and sends it to the server.

On the other hand, once the data acquisition phase is complete, the collected data is filtered by the master unit and sampled again. The refined results are packaged into a unified data corpus. This data is sent to the server when the master device is idle. On the other hand, data transmission may fail on unstable networks. To allow for errors, the master agent keeps all data packages until the network is available.



Time 2018. 09. 10. 15:34:51
Station IP 192.168.1.100
Server IP 165.194.89.50

Sampling Parameter
Sampling Rate 1440 cpd
Resampling Rate 1440 cpd

Data Acquisition Status

Air Temperature	23.5 ℃	CO2 Density	416 ppm
Air Humidity	46 %	Soil Temperature	13.7 ℃
Wind Speed	2.2 m/s	Soil Humidity	61 %
Illuminance	5167 lux	Soil Acidity	6.8 pH

Server List
165.194.89.50

Station List
192.168.1.100
192.168.1.110

Sampling Settings
Sampling Rate
Resampling Rate

Figure 8. GUI of monitoring application program

A real-time system management program is implemented on QT framework. A graphic user interface is presented in **Figure 8**. This program allows users to connect their QT-supported smart devices to the server or to all stations in the system. Users can monitor the refined data corpus in real-time and control the parameters for (re)sampling and filtering.

Users can monitor the data package that is just collected and the status of each station. The left side of the **Figure 8** shows the the latest collected data package and the sampling rate. The ’Next’ button allows to users to monitor information about cut-off frequency of each target data in the same window. Also, in most cases, farmers or researchers use monitoring programs anywhere. Therefore this program is implemented to support for detecting the nearby systems and stations. The configuration window appears to the right side of the **Figure 8**. Users can choose which server to connect to. The station list is then updated to the station connected to that server. Sampling and preprocessing parameters can be changed in the same window too.

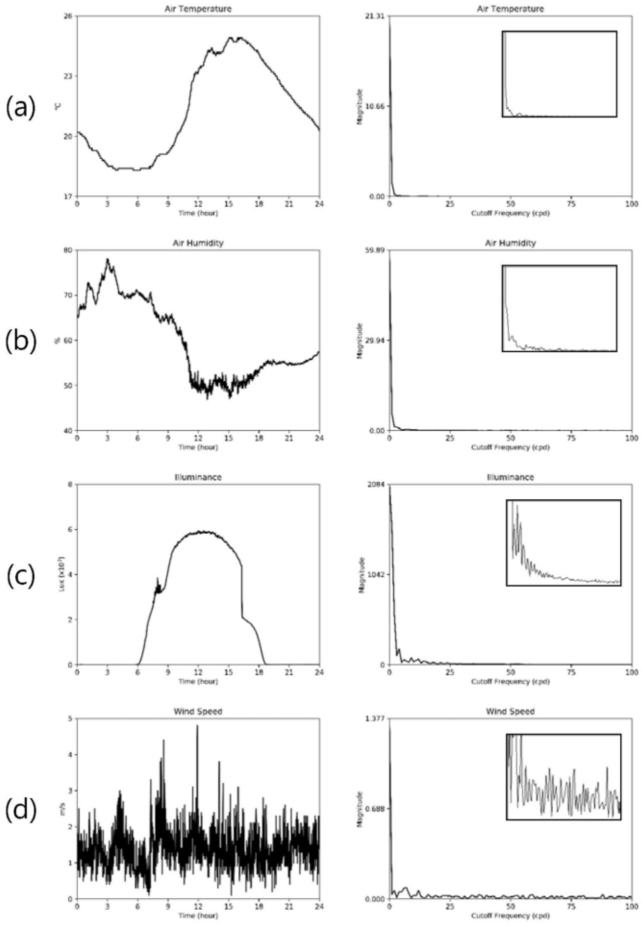


Figure 9. Raw data and spectrum.

3. Experiments

3.1. Data Acquisition

Some of data is selected from **Table 1** and shown as **Figure 9**. Selection is based on their commonly used frequency. In the figure, the time-series data and spectra are placed in the order of temperature, humidity, illuminance and wind speed. The left column is time-series raw data for a day. The other side shows the spectrum for 0∼100 *cpd*. For better understanding, the enlarged magnitudes are presented in each small boxes.

The wind speed data has the biggest changes among those data. (a), (b) and (c) seems to have a small prime frequency as approximately 1∼1.5 *cpd*, but (d) looks like the random. However there is a global trend on the wind speed data.

It can be seen through the frequency graph that the humidity is noisier than the temperature. This can be caused by differences in the performance or operating mechanism between the temperature and the humidity sensors. However, it is clear that the humidity data should be filtered as in the temperature data. The illuminance shapes as noised a half-sine. Especially the illuminance get almost zero before sunrise and after sunset, which is hard to represent with the sum of some sinusoidal signals. The low-threshold LPF will cause a ringing problem, which will distort the RMSE calculation.

3.2. Data Filtering

Because this system adopted daylight-based sampling frequency, almost environmental factors are over-collected. Thus data filtering procedure may be helpful to reduce the entropy of the data. There are some data compaction method such as run-length coding, which is very powerful when the data has many consecutive and duplicated values. However in this system, the filtered data is downsized by resampling due to the computation cost.

Before filtering the data, the proper LPF should be selected to maintain the informative frequency band. Generally this task is performed by the insight of a well-trained data analyst. For example, the temperature or humidity may vary per half or an hour, then the proper filtering frequency should be 24 to 48 *cpd*. By the way, if the ideal cut-off frequency of LPF is underestimated, then the important parts of data will be eliminated.

The implementation of LPF is based on software FFT of which the time complexity is known as $O(n \cdot logn)$. In the experiment, it takes about 100 *msec* for the agent unit to calculate FFT with 1440 samples. This computation cost is feasible to deploy because the data acquisition does not require low latency, even though monitoring function is attached on it. If the computation is performed on the PC, less time will be taken.

Table 3. Selected LPFs based on proposed method

Noise (%)	1	2	3	4	5	6	7	8	9	10
Temperature (*cpd*)	24	24	12	6	6	6	6	6	6	6
Humidity (*cpd*)	240	48	48	24	24	24	24	12	12	12
Illuminance (*cpd*)	360	240	240	120	96	96	48	48	48	48
Wind speed (*cpd*)	720	720	720	720	720	720	720	720	720	720

From left to right columns for experimental purposes, three different LPFs with cutoff frequencies of 6, 24, and 120 *cpd* were applied to the data set in **Figure 10**. These cut-off frequencies were intentionally chosen because it clearly show the difference in the filtered results. The results of the LPFs applied to the temperature, humidity, illuminance and wind speed data are shown in each row. It can be seen that as the cut-off frequency increases, the filtered signal becomes more complex. For example, the result of 6 *cpd* is too simplistic, while 120 *cpd* seems too complex.

On the other hand, the proposed method provides the optimal LPFs as shown in **Table 3**. Calculations for LPF selection are covered on a server. In the experiment, a server equipped with *Intel i5-8500 3.00GHz* CPU and *DDR4 32GB* RAM was used. It takes about 20 minutes to calculate all RMSEs for one cut-off frequency with [0,50]% ranged noise in steps of 1%. In the experiment, only several cut-off frequencies: 3, 6, 12, 24, 48, 96, 120, 240, 360 are selected to reduce the processing time. it takes about 3 hours for calculation on the server.

The table shows the threshold changes according to noise ratio. The larger the noise, the lower the cut-off frequency is adopted. However, this method is not effective for wind

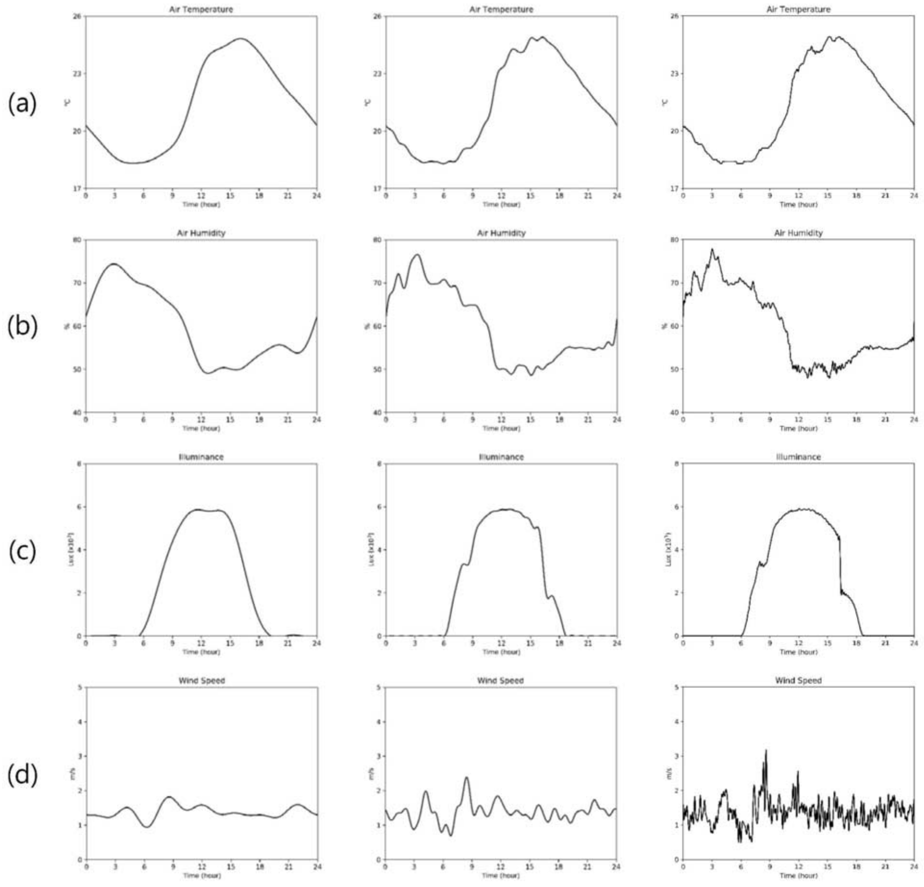


Figure 10. Filtered signal

speed because wind speed data is also distributed in the high frequency band. Therefore, even the RMSE between the source data and the filtered data is minimized, so the result is not plausible in the case of wind speed. As a result, if the period of data change is so short that the frequency components get high, it is recommended to choose the LPF in manual or to use the raw data directly.

The results of RMSE are compared for further analysis in **Figure 11**. The RMSE between original and filtered signal is compared according to cut-off frequency changes. For each data type, the RMSE is normalized based on the value of the LPF with 1 *cpd*. This graph shows the effects of the LPF. As the cut-off frequency approaches 1 *cpd*, the RMSE increases significantly indicating significant loss of information. In contrast, the RMSE decreases slightly from a certain point as the cut-off frequency increases. These inflection points can serve as a basis for judging whether significant data is retained.

Important observations are also graphical gradients. Temperature, humidity, and illuminance data appear as a fractional function. This means that there is a certain threshold that distinguishes the genuine data from the noise. It is clear that the noise can be

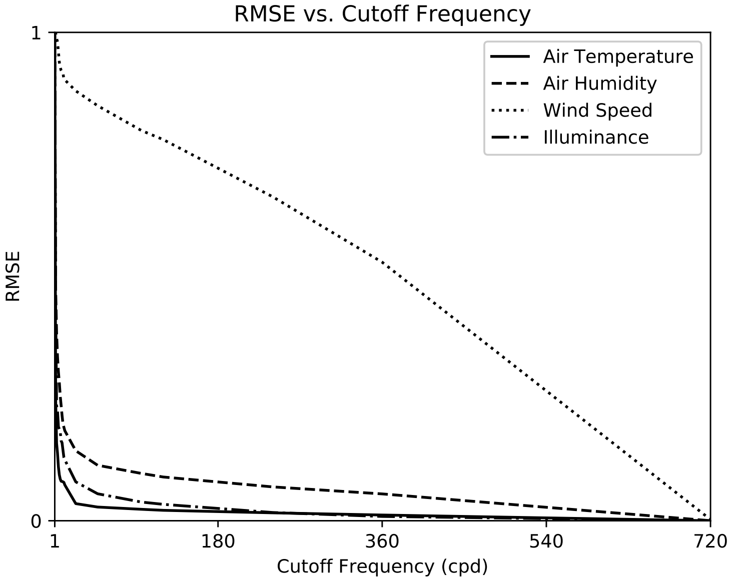


Figure 11. RMSE comparison according to cut-off frequency changes

easily separated as the shape of the graph approaches the two axes. However, the wind speed is almost linear. It proves the informative data is spread over entire frequency band.

The data type also has a lower cut-off frequency in the order close to the origin of the coordinate, excluding illuminance. This trait helps decision-making or automation in data processing. On the other hand, since the illuminance graph shape as a half-sine, the fundamental frequency should be the smallest, but it is not. This indicates that the RMSE of the illumination data is distorted due to the ringing problem. In this case, this problem can be solved by selectively filtering only when the sun comes up.

3.3. Noise Endurance

Noise interference is inevitable during the measurement and signal digitization. Several simulations were performed to analyse the noise immunity characteristics. A random noise was inserted to the raw data and the LPFs of which cut-off frequencies are 3, 12, 48, 120 and 360 *cpd* each were then applied in the simulations. The reason why these cut-off frequencies are selected is because they are clearly separated in the graph so that easy to understand. Random noises between 0% to 50% are synthesized with the original raw data to verify the effect of each LPFs, which is represented as the RMSE.

As shown in the **Figures 12** and **13**, the relationships among three variables: noise ratio, cut-off frequency and effect of LPF, are clearly compared. All RMSEs are normalized to provide better understanding. Original graph is not filtered and can be used as an indicator that shows the threshold effect of LPF. For example, every LPFs where their RMSEs are larger than the original one are regarded as not effective. Considering all LPFs has an exponential shape, each LPF have their own turning point where filtering becomes effective.

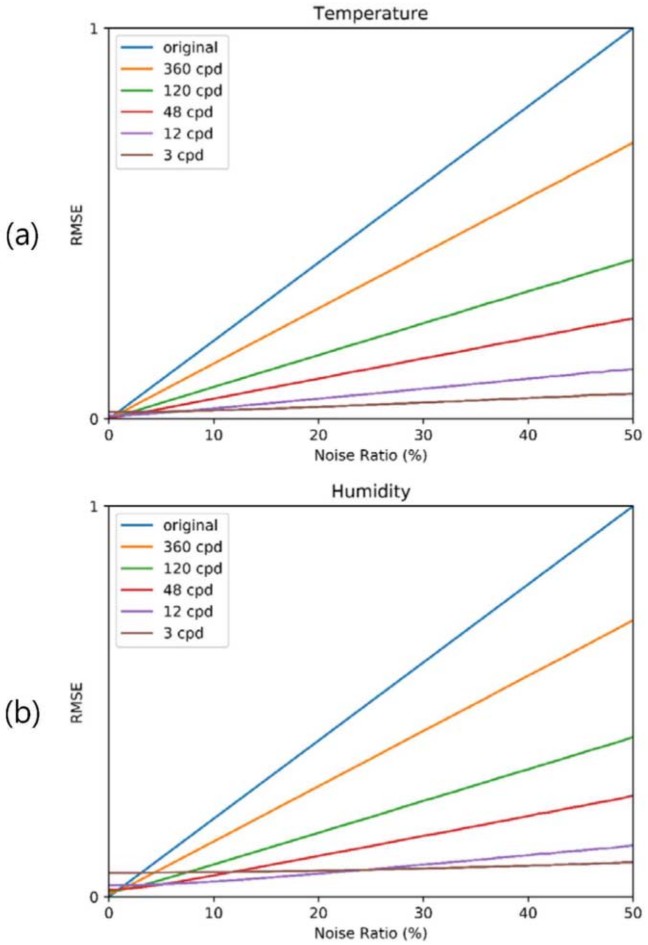


Figure 12. RMSE comparison of (a) temperature and (b) humidity data according to noise ratio changes

On the other hand, the proposed LPF selection method can be easily explained in the figure. For given noise ratio $p\%$, a vertical line can be drawn on that. Then the cross-points between LPFs and the line are the expected effect of LPF on given noise ratio $p\%$. Among them, the LPF that has the least RMSE is revealed as the optimal filter.

Figure 12 tells the frequency distribution of temperature and humidity data is concentrated at low frequencies. This can be found from the RMSE gap between 3 *cpd* and 360 *cpd*. If there is a significant gap between any LPFs, that cut-off frequency range might be the primary band of original data. For example, **Figure 12**-(b) shows the informative frequency band is larger than **Figure 12**-(a) because the RMSE gap in (b) is bigger than (a).

Moreover it can be determined whether or not genuine data has been lost significantly based on the intersection point. If the noise ratio is greater than the crossover section, noise can be reduced no matter what cut-off frequency is used. On the other hand, where the noise ratio is less than the intersection point, some low cut-off frequencies

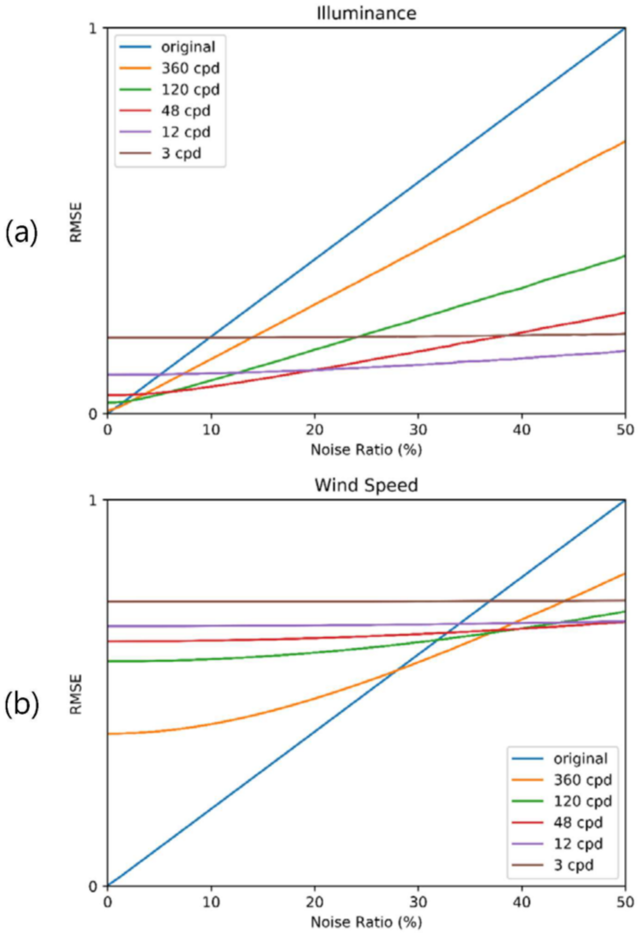


Figure 13. RMSE comparison of (a) temperature and (b) humidity data according to noise ratio changes

are shown in the figure that significantly corrupt the informative data. However, when using these filters, strong noise immunity can be attained, as seen in a gradual slope in the picture.

The illuminance and wind speed have much higher primary band than the one of humidity as shown in **Figure 13**. The wind speed data seems not to need any filter with under 30% noise. It results from the frequency characteristic of the wind speed that the meaningful data is distributed over the spectrum.

On the other hand, the RMSE of illuminance with lower cut-off frequency seems to be overestimated because of the ringing problem. Unlike the other data, the RMSE gap between 48 *cpd* and 120 *cpd* is slightly smaller than the between 120 *cpd* and 360 *cpd*. It reveals that there is a constant value interval in the data which occurs the ringing problem. As mentioned, however, it can be solved by various selective filtering method. Thus it is not considerably concerned in this paper.

4. Summary and Future Works

A novel data acquisition and preprocess system for precision agriculture is presented. Various heterogeneous data stream can be easily attained with user-friendly device. All attached sensors have the same sampling period based on daylight, which was proved as being proper. Proposed system is fully implemented and deployed in controlled outdoor environment. RMSE-based preprocessing method is also proposed and the comparison and analysis are held. This system automatically generates the optimal LPF with given expecting noise ratio.

This system can provide a well-refined heterogeneous agricultural dataset on the farmland. Even though farmer does not have any electronic knowledges, he or she can easily attach sensor because this system is based on ready-made device, *RaspberryPi*. Any image devices for phenotypic purpose can be attached in this system because proposed system is based on Linux OS. However this system does not equipped with any error endurance or error pruning supports. Thus when deploying this system into a real farmland, additional system support should be adopted such as reusable power, counterplan for rain and so on.

On the other hand, gathered data will form a big data, which is enable to in-depth analysis. Lots of applications using several information technologies will be flourished with those data [27,28].

Acknowledgements

This research was supported by Basic Science Research Programs through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2016R1D1A1B03936349) and by the Business for Cooperative R&D between Industry, Academy, and Research Institute through the Korea Small and Medium Business Administration in 2016 under Grant C0443994.

References

[1] J. Zhao and J. Tang, Understanding agricultural growth in China: An international perspective, *Structural Change and Economic Dynamics* **46** (2018), 43–51, ISSN 0954-349X. doi:10.1016/j.strueco.2018.03.006.

[2] D.J. Mulla, Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps, *Biosystems Engineering* **114**(4) (2013), 358–371, Special Issue: Sensing Technologies for Sustainable Agriculture, ISSN 1537-5110. doi:10.1016/j.biosystemseng.2012.08.009.

[3] J.W. Jones, J.M. Antle, B. Basso, K.J. Boote, R.T. Conant, I. Foster, H.C.J. Godfray, M. Herrero, R.E. Howitt, S. Janssen, B.A. Keating, R. Munoz-Carpena, C.H. Porter, C. Rosenzweig and T.R. Wheeler, Brief history of agricultural systems modeling, *Agricultural Systems* **155** (2017), 240–254, ISSN 0308-521X. doi:10.1016/j.agsy.2016.05.014.

[4] X. Pham and M. Stack, How data analytics is transforming agriculture, *Business Horizons* **61**(1) (2018), 125–133, ISSN 0007-6813. doi:10.1016/j.bushor.2017.09.011.

[5] A. Tzounis, N. Katsoulas, T. Bartzanas and C. Kittas, Internet of Things in agriculture, recent advances and future challenges, *Biosystems Engineering* **164** (2017), 31–48, ISSN 1537-5110. doi:10.1016/j.biosystemseng.2017.09.007.

[6] J.M. Antle, B. Basso, R.T. Conant, H.C.J. Godfray, J.W. Jones, M. Herrero, R.E. Howitt, B.A. Keating, R. Munoz-Carpena, C. Rosenzweig, P. Tittonell and T.R. Wheeler, Towards a new generation of agricul-

tural system data, models and knowledge products: Design and improvement, *Agricultural Systems* **155** (2017), 255–268, ISSN 0308-521X. doi:10.1016/j.agsy.2016.10.002.

[7] D. Chen, K. Neumann, S. Friedel, B. Kilian, M. Chen, T. Altmann and C. Klukas, Dissecting the Phenotypic Components of Crop Plant Growth and Drought Responses Based on High-Throughput Image Analysis, *The Plant Cell* **26**(12) (2014), 4636–4655, ISSN 1040-4651. doi:10.1105/tpc.114.129601.

[8] S. Wolfert, L. Ge, C. Verdouw and M.-J. Bogaardt, Big Data in Smart Farming – A review, *Agricultural Systems* **153** (2017), 69–80, ISSN 0308-521X. doi:10.1016/j.agsy.2017.01.023.

[9] S. Singh, M. Bergerman, J. Cannons, B. Grocholsky, B. Hamner, G. Holguin, L. Hull, V. Jones, G. Kantor, H. Koselka, G. Li, J. Owen, J. Park, W. Shi and J. Teza, Comprehensive Automation for Specialty Crops: Year 1 results and lessons learned, *Intelligent Service Robotics* **3**(4) (2010), 245–262, ISSN 1861-2784. doi:10.1007/s11370-010-0074-3.

[10] J.L. Araus and J.E. Cairns, Field high-throughput phenotyping: the new crop breeding frontier, *Trends in Plant Science* **19**(1) (2014), 52–61, ISSN 1360-1385. doi:10.1016/j.tplants.2013.09.008.

[11] J.W. White, P. Andrade-Sanchez, M.A. Gore, K.F. Bronson, T.A. Coffelt, M.M. Conley, K.A. Feldmann, A.N. French, J.T. Heun, D.J. Hunsaker, M.A. Jenks, B.A. Kimball, R.L. Roth, R.J. Strand, K.R. Thorp, G.W. Wall and G. Wang, Field-based phenomics for plant genetics research, *Field Crops Research* **133** (2012), 101–112, ISSN 0378-4290. doi:10.1016/j.fcr.2012.04.003.

[12] F.A. van Eeuwijk, D. Bustos-Korts, E.J. Millet, M.P. Boer, W. Kruijer, A. Thompson, M. Malosetti, H. Iwata, R. Quiroz, C. Kuppe, O. Muller, K.N. Blazakis, K. Yu, F. Tardieu and S.C. Chapman, Modelling strategies for assessing and increasing the effectiveness of new phenotyping techniques in plant breeding, *Plant Science* (2018), In Press, ISSN 0168-9452. doi:10.1016/j.plantsci.2018.06.018.

[13] S.C. Gosa, Y. Lupo and M. Moshelion, Quantitative and comparative analysis of whole-plant performance for functional physiological traits phenotyping: New tools to support pre-breeding and plant stress physiology studies, *Plant Science* (2018), In Press, ISSN 0168-9452. doi:10.1016/j.plantsci.2018.05.008.

[14] J.M. Talavera, L.E. Tobón, J.A. Gómez, M.A. Culman, J.M. Aranda, D.T. Parra, L.A. Quiroz, A. Hoyos and L.E. Garreta, Review of IoT applications in agro-industrial and environmental fields, *Computers and Electronics in Agriculture* **142** (2017), 283–297, ISSN 0168-1699. doi:10.1016/j.compag.2017.09.015.

[15] M. Ruiz-Altisent, L. Ruiz-Garcia, G.P. Moreda, R. Lu, N. Hernandez-Sanchez, E.C. Correa, B. Diezma, B. Nicolaï and J. García-Ramos, Sensors for product characterization and quality of specialty crops—A review, *Computers and Electronics in Agriculture* **74**(2) (2010), 176–194, ISSN 0168-1699. doi:10.1016/j.compag.2010.07.002.

[16] Y. Huang, Z.-x. CHEN, T. YU, X.-z. HUANG and X.-f. GU, Agricultural remote sensing big data: Management and applications, *Journal of Integrative Agriculture* **17**(9) (2018), 1915–1931, ISSN 2095-3119. doi:10.1016/S2095-3119(17)61859-8.

[17] W.S. Lee, V. Alchanatis, C. Yang, M. Hirafuji, D. Moshou and C. Li, Sensing technologies for precision specialty crop production, *Computers and Electronics in Agriculture* **74**(1) (2010), 2–33, ISSN 0168-1699. doi:10.1016/j.compag.2010.08.005.

[18] A. Kamilaris, A. Kartakoullis and F.X. Prenafeta-Boldú, A review on the practice of big data analysis in agriculture, *Computers and Electronics in Agriculture* **143** (2017), 23–37, ISSN 0168-1699. doi:10.1016/j.compag.2017.09.037.

[19] F. Al-Turjman, The road towards plant phenotyping via WSNs: An overview, *Computers and Electronics in Agriculture* (2018), In Press, ISSN 0168-1699. doi:10.1016/j.compag.2018.09.018.

[20] F.J. Mesas-Carrascosa, D.V. Santano, J.E. Meroño, M.S. de la Orden and A. García-Ferrer, Open source hardware to monitor environmental parameters in precision agriculture, *Biosystems Engineering* **137** (2015), 73–83, ISSN 1537-5110. doi:10.1016/j.biosystemseng.2015.07.005.

[21] R. Aquino-Santos, A. González-Potes, A. Edwards-Block and R.A. Virgen-Ortiz, Developing a New Wireless Sensor Network Platform and Its Application in Precision Agriculture, *Sensors* **11**(1) (2011), 1192–1211, ISSN 1424-8220. doi:10.3390/s110101192.

[22] G. Bai, Y. Ge, W. Hussain, P.S. Baenziger and G. Graef, A multi-sensor system for high throughput field phenotyping in soybean and wheat breeding, *Computers and Electronics in Agriculture* **128** (2016), 181–192, ISSN 0168-1699. doi:10.1016/j.compag.2016.08.021.

[23] D. Reynolds, F. Baret, C. Welcker, A. Bostrom, J. Ball, F. Cellini, A. Lorence, A. Chawade, M. Khafif, K. Noshita, M. Mueller-Linow, J. Zhou and F. Tardieu, What is cost-efficient phenotyping? Optimizing costs for different scenarios, *Plant Science* (2018), In Press, ISSN 0168-9452. doi:10.1016/j.plantsci.2018.06.015.

[24] S.J.C. Janssen, C.H. Porter, A.D. Moore, I.N. Athanasiadis, I. Foster, J.W. Jones and J.M. Antle, Towards a new generation of agricultural system data, models and knowledge products: Information and communication technology, *Agricultural Systems* **155** (2017), 200–212, ISSN 0308-521X. doi:10.1016/j.agsy.2016.09.017.

[25] S. Fritz, L. See, J.C.L. Bayas, F. Waldner, D. Jacques, I. Becker-Reshef, A. Whitcraft, B. Baruth, R. Bonifacio, J. Crutchfield, F. Rembold, O. Rojas, A. Schucknecht, M.V. der Velde, J. Verdin, B. Wu, N. Yan, L. You, S. Gilliams, S. Mücher, R. Tetrault, I. Moorthy and I. McCallum, A comparison of global agricultural monitoring systems and current gaps, *Agricultural Systems* **168** (2019), 258–272, ISSN 0308-521X. doi:10.1016/j.agsy.2018.05.010.

[26] A.G. Maia, B.C.B. Miyamoto and J.R. Garcia, Climate Change and Agriculture: Do Environmental Preservation and Ecosystem Services Matter?, *Ecological Economics* **152** (2018), 27–39, ISSN 0921-8009. doi:10.1016/j.ecolecon.2018.05.013.

[27] R.H.L. Ip, L.-M. Ang, K.P. Seng, J.C. Broster and J.E. Pratley, Big data and machine learning for crop protection, *Computers and Electronics in Agriculture* **151** (2018), 376–383, ISSN 0168-1699. doi:10.1016/j.compag.2018.06.008.

[28] A. Kamilaris and F.X. Prenafeta-Boldú, Deep learning in agriculture: A survey, *Computers and Electronics in Agriculture* **147** (2018), 70–90, ISSN 0168-1699. doi:10.1016/j.compag.2018.02.016.

Agriculture and Environment Perspectives in Intelligent Systems
A. Muñoz and J. Park (Eds.)
IOS Press, 2019

doi:10.3233/AISE190005

Prediction of Uptake of Carbamazepine and Diclofenac in Reclaimed Water-Irrigated Lettuces by Machine Learning Techniques

Raquel Martínez-España[a,1], Andrés Bueno-Crespo[a], Mariano González García[b] and Carmen Fernández-López[c]

[a] *Computer Department, Catholic University of Murcia (UCAM), Murcia. Spain.*
[b] *International University of La Rioja (UNIR), La Rioja, Spain*
[c] *University Centre of Defense at the Spanish Air Force Academy, Santiago de la Ribera, Spain*

Abstract. Currently, due to the global shortage of water, the use of reclaimed water from the Wastewater Treatment Plants (WWTPs) for the irrigation of crops is an alternative in areas with water scarcity. However, the use of this reclaimed water for vegetable irrigation is a potential entry of pharmaceutical products into the food chain due to the absorption and accumulation of these contaminants in different parts of the plants. In this work we carried out an analysis of five machine learning techniques (Random Forest, support vector machine, M5 Rules, Gaussian Process and artificial neural network) to predict the uptake of carbamazepine and diclofenac in reclaimed water-irrigated lettuces with the consequent saving of environmental and economic costs. For the different combinations of input and output, the prediction results using the of machine learning techniques proposed on the pharmaceutical components in reclaimed water-irrigated lettuces are satisfactory, being the best technique the Random Forest that obtains a model fit value (R^2) higher than 96.5% using a single input in the model and higher than 97% using two inputs in the model.

Keywords. machine learning, carbamazepine, diclofenac, reclaimed water-irrigated, lettuces

1. Introduction

Over the years, many advanced societies have found uses for reclaimed, or recycled, water to both conserve fresh water supplies and improve economic growth and development. While some drought prone and arid regions have used reclaimed water since the 1920s, more cities and regions worldwide are now building water reclamation into their planning practices.

[1] Raquel Martínez-España, Computer Department, Catholic University of Murcia (UCAM), Campus de los Jerónimos, Guadalupe 30107, Murcia. Spain.; E-mail: rmartinez@ucam.edu.

There are many kinds of water, each of which falls under the category of “drinking water” (potable) or “non-potable” (not suitable for consumption). In the U.S., tap water and well water are usually the only types of water that are considered drinking water. Nevertheless, well water is not regulated by the Environmental Protection Agency (EPA) [1], so some well water might not be drinking water. Non potable types include reclaimed water, salty water, sea water, river water and runoff.

Reclaimed water is basically “cleaned” wastewater. It’s the final product of a multi-stage, advanced water treatment process. This processed water is normally focus on severe standards and is severely monitored by local, regional and state governments to ensure it continuously meets those standards.

Occurrence and fate of pharmaceuticals and personal care products (PPCPs) in the environment and the potential consequences for human health is of serious concern. Conventional wastewater treatment plants (WWTPs) are only moderately effective at removing many organic contaminants, including PPCPs [2, 3].

It is thus necessary to assess the behavior of these compounds. Hardly any studies have been followed through with measured concentrations of these PPCPs in the wastewater influent and effluent in WWTPs throughout the Murcia Region [4]. The Region of Murcia is located in southeastern Spain (Iberian Peninsula). It has an extension of 11,313 km2 and is entirely within the Segura River Basin [5]. The climate of this region is semi-arid Mediterranean, with an average annual temperature of 18 ºC and scarce annual rainfall, around 300 mm [6]. In spite of the scarce natural water resources, there is sturdy use of water for urban, industrial and irrigation purposes [7].

The analysis of concentrations of these PPCPs has high economic and environmental costs. In addition to these costs, the slowness of these analyses must be taken into account if large quantities of effluents from WWTPs are to be analysed. Therefore, to address this problem, in this study we propose the use of machine learning techniques to predict the absorption of concentrations of these PPCPs, specifically carbamazepine (CBZ) and diclofenac (DCF) in lettuce irrigated by effluents from WWTPs. Machine learning techniques allow us to build computational models that learn through data and enabling to create behavioral patterns that serve as the basis for decision making. In this work we propose the study of five machine learning techniques to find the best model to predict the absorption of carbamazepine and diclofenac in the leaves and roots of lettuce. Specifically, we will use the Random Forest technique, the artificial neural network, the M5 rules technique, the Gaussian Process technique and the support vectorial machine. For this study we will carry out different experiments taking as input to the different inputs techniques seeking to predict with the best possible adjustment the concentration of carbamazepine and diclofenac in the water, as well as the absorption of these compounds in the leaves and roots of lettuce.

The main goal of this study is to analyze the set of machine learning techniques discussed to determine the best model for predicting CBZ and DFC in the leaves and roots of lettuce irrigated by effluents from WWTPs. In this way, environmental and economic savings are achieved in laboratory analyses.

The study is structured as follows. In section 2 a background analysis of the importance of detecting residues of compounds in the food chain is carried out. Section 3 describes the data collection methodology used for this study and a description of the machine learning techniques used is presented. In Section 4 the experiments carried out with machine learning techniques are described, analyzed and discussed. Finally, section 5 presents the conclusions and future work.

2. Background

In countries with water scarcity, it is necessary to use non-traditional water resources, such as treated wastewater to irrigate crops. However, this type of irrigation presents potential risks because water often contains emerging contaminants such as pharmaceutical compounds (PCs) that are absorbed by the roots and later transport and accumulate in the edible portion (the leaves) of the vegetables (Figure 1). Consequently, this make available a means of route of the PCs into the food chain [8 ,9, 10, 11]. Within these PCs we find the CBZ and the DCF.

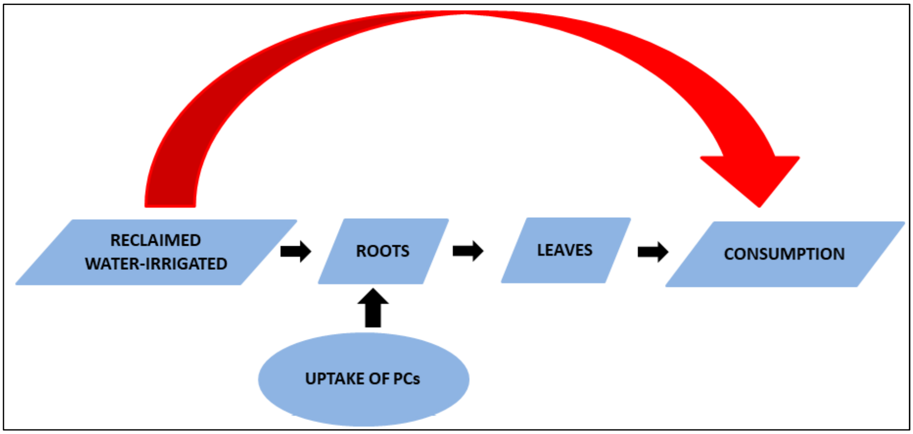


Figure 1. Process of absorption and transport of PCs in vegetables tissues.

Carbamazepine (CBZ) is a PC that is used to treat several medical conditions such as epilepsy [12]. CBZ is metabolized in the human body and approximately 30% reaches the environment through wastewater due to its low elimination efficiency in wastewater treatment plants [13, 14]. Diclofenac is a PC commonly used as an analgesic and antiarthritic [15] and toxicological effects have been detected in several organisms such as fish and mussels at low levels [16].

Several studies have been carried out on the absorption of several PCs to confirm the accumulation of certain compounds in the tissues of plants [17], for example, Shenker et al. (2011) [18] demonstrated an uptake of CBZ by cucumber plants in hydroponic crops and experiments in greenhouses. Another study reported on the absorption and transformation of the CBZ into tomato plants [19]. In the case of diclofenac, there are numerous studies conducted such as an absorption study was conducted comparing the accumulation of four common PCs, bisphenol A (BPA), diclofenac (DCF), naproxen (NPX) and nonylphenol (NP), in lettuce and cabbage, examining the distribution of the accumulated PCs [20].

Due to the different mechanisms of PC accumulation in crops that are irrigated with recovered water, it is essential to continue investigating PC accumulation in lettuce in the Region of Murcia, where the use of recovered water is very important because it has an area of 188,000 hectares that correspond to agricultural land, in which consumption (urban, industrial and agricultural) has increased in recent years, which has resulted in overexploitation of water [21].

3. Material and Methods

3.1. Experiment description

The data used to validate the expert system proposed in this study are taken from a real scenario as described in González García et al. (2018) [22]. Briefly, three varieties of lettuce (Lactuca sativa) were grown in a greenhouse with controlled different parameters. Plants were grown in bags of coconut fibre substrate and drip irrigated with effluent from a WWTP. Figure 2 shows a representation of the WWTP studied. The WWTP receives influent and the wastewater goes through a primary treatment (1) consisting of a screen, a grit removal tank and a primary clarifier (2). The wastewater is then biologically treated (3), secondary treatment (4) and filtered through a layer of a sand (5) before finally being disinfected with ultraviolet radiation (6).

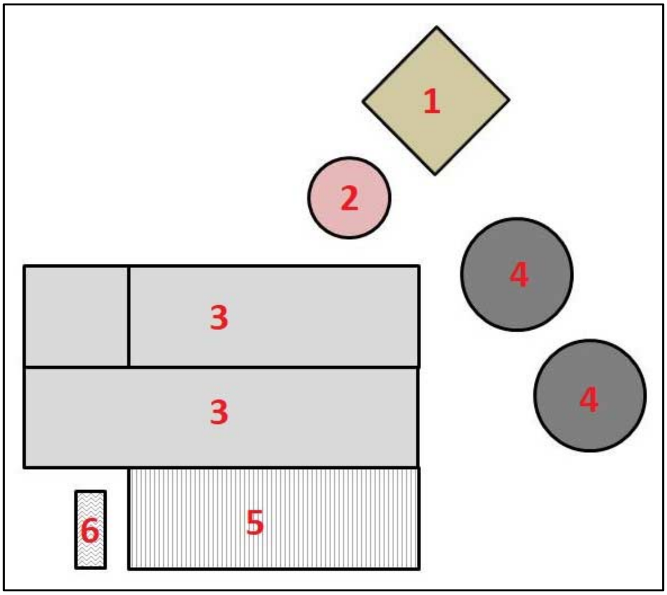


Figure 2. Representation of the WWTP studied.

The plants were provided for nutrients dissolved in water containing different concentrations of CBZ and DCF. Treatments consisted of the application of 0, 30, 60, 120 and 210 μg•L-1 of CBZ and DCF for a period of two weeks. After two weeks of treatment, plant roots and leaves were separated and analyzed with ultra-performance liquid chromatography (UPLC). This technique is used in analytical chemistry to separate, identify and quantify each component in a mixture. Acquity UPLC I-Class System and HR-QTOF-MS maXis Series (Daltonik GmbH, Bruker, Germany) instruments were used. Roots and leaves were cleaned with deionized water before preparing the UPLC samples. Each plant compartment was freeze-dried and weighed to determine its dry mass. The material was then ground to a fine powder. Sample extraction was performed according to a method reported elsewhere [23, 24].

In brief, a 0.2 g aliquot of freeze-dried plant material was extracted in an ultrasonic water bath for 20 min using 20 mL of methyl tert-butyl ether (MTBE) as the solvent. This was followed by centrifugation at 3,000 rpm for 20 min. The supernatant was decanted into a 40 mL glass vial and the residue extracted using 20 mL of acetonitrile. The aqueous solution was loaded onto a hydrophilic–lipophilic balance cartridge and preconditioned. The sample was loaded onto the cartridges at a rate of 15 mL/min and the cartridges were then eluted into 15 mL calibrated centrifuge tubes. The resulting extract was concentrated to dryness and brought to a final volume of 1 mL using methanol. Samples were filtered using nylon filters (25 mm, 0.45 μm pore size). Each sample was prepared in duplicate. An outline of this described process is shown in Figure 3.

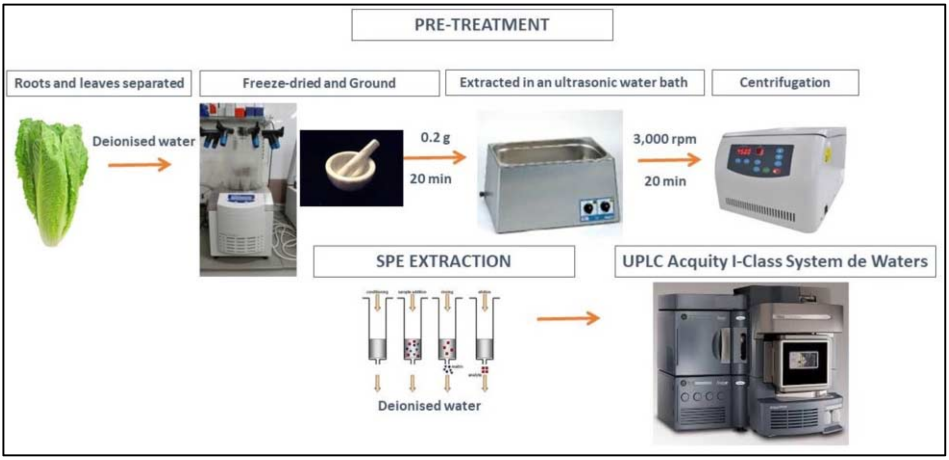


Figure 3. Diagram of the sample preparation technique and laboratory analyses.

3.2. Description of the methods of machine learning

This section shows the experiments that have been addressed in this work. For this, five supervised machine learning techniques have been compared; regression rules (M5R), random forests (RF), Bayesian Gaussian (GP), Support Vector Machine (SVM) and Artificial Neural Network (ANN). These five techniques have been selected as a representation of the different classifications that can be found within the supervised techniques. Below is a brief description of each of the techniques. Then, in the experiments section, the most significant results are shown (note that all the results are shown in an appendix), and finally the results are discussed.

- M5R: This technique is known as regression rules M5Rules. It is a regression rule technique that uses a system of separate-and-conquer and that is based on classic tree models [25, 26]. In an iterative process, trees are generated, from which the best sheet is selected to build a rule. This technique uses exceptions with respect to the rule that has been generated to create rules that reduce the error.

- RF: This technique constructs random forests using bagging ensembles of random trees [27]. For the construction of the decision trees, in this work the algorithm C4.5 has been used, which is based on a pruning system combined with the gain of information. It allows working with lost values and although in this work it is used for regression, it is also a technique widely used in regression problems [28].

- GP: This technique uses Bayesian Gaussian in regression problems, specifically when the problem is not linear [29]. This technique allows choosing the training data to be normalized and requires two parameters to perform the regression process, on the one hand, a kernel function and a noise regularization parameter [30].

- SVM: This technique is both a supervised learning method for two classes where a map of the data is generated, creating a margin of separation between them, as much as possible [31]. Along with artificial neural networks, SVMs are used in many applications for both classification and regression. Although they allow working with linear and non-linear models, done in the linear models a function (kernel) is used to increase the dimensionality of the model, in this work we have used a linear model, specifically a polynomial kernel since it is the one that better fits the problem [32].

- ANN: This technique is inspired by biological neural networks. An artificial neural network (ANN) is composed of a large number of nodes known as artificial neurons (perceptron) interconnected between them, so that the learning process is done through examples [33, 34]. This learning process simulates the behavior of the brain where neurons, interconnected between them, transmit a signal from one another, producing the learning process by adjusting those signals (known as synapses for the biological neuron and as a weight for the artificial one). In this work, a trained multilayer perceptron (MLP) with backpropagation has been used [35].

4. Experiments

In this section we are going to detail a set of experiments that we have carried out in order to evaluate the different machine learning techniques and analyse the best configuration of input and output inputs in order to avoid laboratory analyses, thus saving both economic costs and environmental costs. In order to carry out these experiments we have used the data set described in section 3.1. The variables that the dataset has available are the concentration in the water of the drug (Conc), the absorption in the leaf of the lettuce of the drug (Leaf), the absorption in the root of the lettuce of the drug (Root) and cultivar (Cult). The validation method used is a 3 fold cross validation. In addition we have repeated each experiment 5 times to evaluate the quality of the models obtained and the stability and robustness of the techniques evaluated.

For both the CBZ and the DCF, we are going to carry out a study carrying out different combinations between the input and output elements, in order to find out in case of having to carry out a laboratory analysis which element is better to analyse in order to obtain a better prediction by the techniques. Thus, we are going to introduce a single input variable or introduce two input variables to each of the machine learning techniques that we have presented in order to find the best model taking into account the economic and environmental savings and looking for the best precision and the best quality in the models. The measures used to analyze and study the results obtained are:

- The correlation coefficient of Pearson (r) that indicates the statistical correlation between the value predicted by the technique and the real value. Value between -1 and 1 the closer to the extremes the better the correlation between the variables.
- The coefficient of determination (R^2) that indicates the goodness of adjustment of the data to the constructed model. Value between 0 and 1, the closer to 1 the better the fit of the model.
- Standard deviation (STD) that indicates the variance that suffers the adjustment of the model according to the input data, that is to say, it indicates us if there is dependence or not of the models with the data. The smaller the value, the more robust the technique, since it has less variance.

4.1. Results

The results are presented in two stages, in the first, all the combinations of experiments have been made where for the different combinations of inputs, outputs and the five methods of machine learning, the correlation coefficient of Pearson, the coefficient of determination (R^2) and its standard deviation are shown. All the experiments have been performed for both the CBZ and the DCF. Specifically, Table 3 shows in detail the results for the CBZ prediction taking a single variable as the input for the technique. Table 4 shows the results for CBZ prediction using two input variables. Table 5 shows the results obtained for the DCF prediction considering a single variable as an input to the techniques and Table 6 shows the DCF prediction using two input variables. These four tables are found in the appendix.

Figures 4, 5, 6 and 7 show the different methods related to the input and output of CBZ y DCF. In the legend of the figures we have separated by '-' the input variables of the output variables and separated by '/' the different input variables. For example, when we have a single input variable in the legend, "Conc-Root" is specified, which means that as an input to the technique, the concentration in the water is taken and as output we expect to predict the uptake of the pharmaceutical remaining in the root. The same happens when we have two variables as input, in this case "Conc/Leaf-Root" means that the technique takes as input the concentration of the pharmaceutical in the water and the absorption of the pharmaceutical in the leaf and as output predicts the absorption of the pharmaceutical in the root. Before analyzing in detail all the results it is important to mention that in general the models obtained by all the techniques obtain a good fit and a fairly low deviation, which reflects the robustness of the models built. Therefore we can affirm that both using one input variable and two input variables the results are satisfactory.

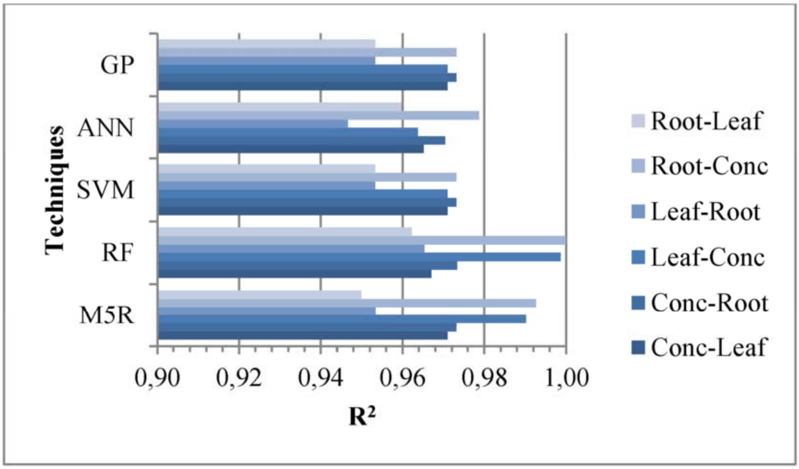


Figure 4. Carbamacepine results from a single input variable

Figure 4 shows the results R^2 of the five machine learning techniques for predicting CBZ using a single input variable. In Figure 5, two inputs have been used. It is observed that when the input is a leaf or a root and it is intended to calculate the root and the leaf the selected method is less reliable than when the input variables that are involved are the concentration. The best models are obtained by taking the water CBZ concentration as output. The best result for both one entry and two entries is always obtained by the Random Forest technique and always predicting the concentration of the pharmaceutical compound in the water. The technique that obtains the worst models globally is the ANN.

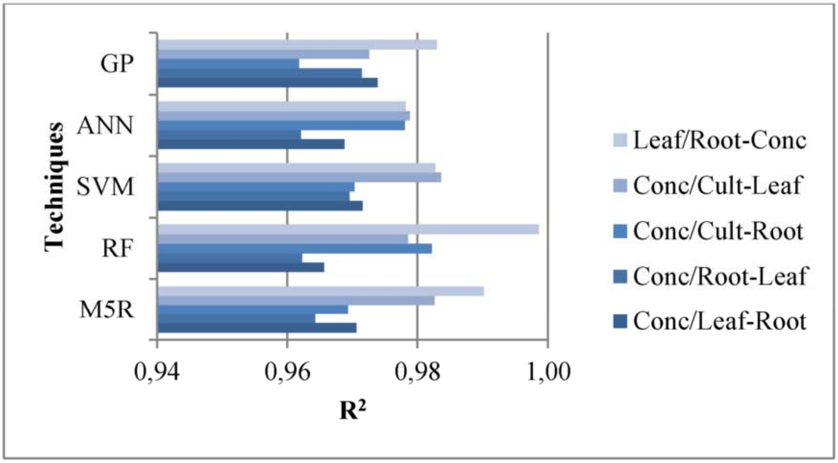


Figure 5. Carbacepine results from two input variable

Figures 6 and 7 show the results of R^2 for DCF considering one and two input variables respectively. As with CBZ, when predicting the concentration of DCF in water, the best results are obtained, with the best technique being the Random Forest assembly. We cannot emphasize a technique that obtains worse models than another, since the behavior of the five techniques is very similar. Although the result obtained by the GP technique as it gets a little worse models for both an entry and for two entries globally.

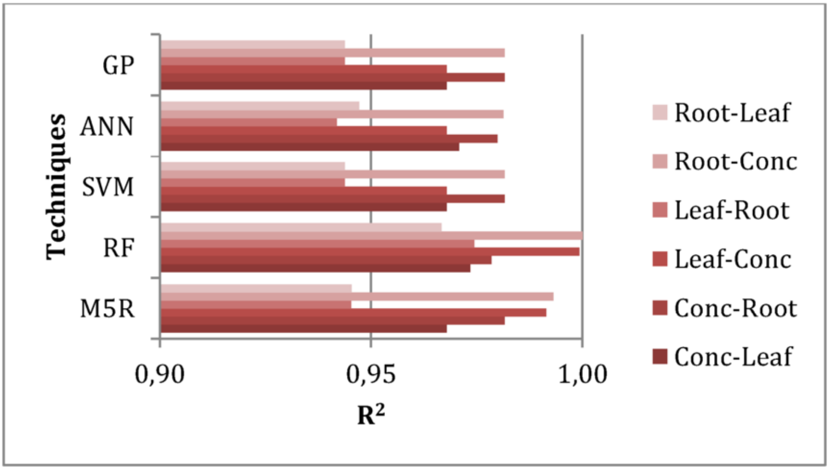


Figure 6. Diclofenac results from a single input variable

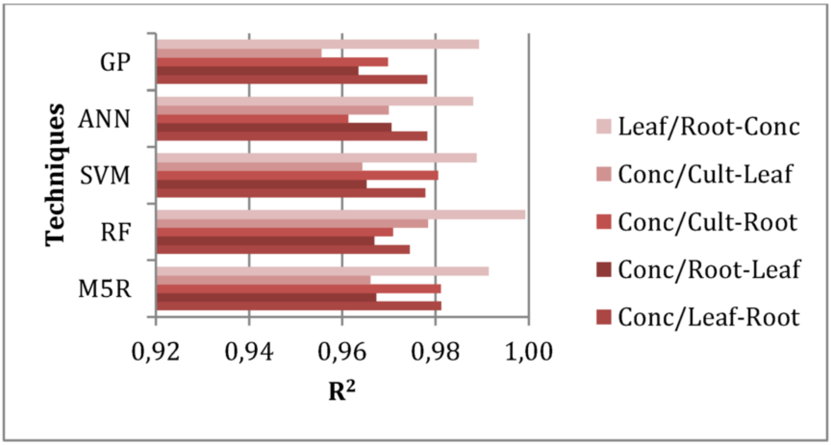


Figure 7. Diclofenac results from two input variable

Once the results have been analysed individually between the CBZ and DCF drugs, we are going to draw up a summary table to analyse the results of both drugs together and highlight the best technique for each of the possible inputs and outputs. Tables 1 and 2 show the summary for each combination of input, output and pharmaceutical compound (DFC and CBZ), indicating the best method is shown according to R^2. The difference between Table 1 and 2, is determined by the input, since in Table 1 a single input is used, in the second one, two inputs are used.

Table 1. Best method for each component for a single input.

Input	Output	Compound	Method	R^2
Concentration	Leaves	CBZ	SVM and M5Rules	0,9710
		DFC	RandomForest	0.9735
Concentration	Root	CBZ	RandomForest	0.9733
		DFC	SVM and M5Rules	0.9816
Leaves	Concentration	CBZ	RandomForest	0.9986
		DFC	RandomForest	0.9993
Leaves	Root	CBZ	RandomForest	0.9653
		DFC	RandomForest	0.9744
Root	Concentration	CBZ	RandomForest	0.9999
		DFC	RandomForest	1..000
Root	Leaves	CBZ	RandomForest	0.9622
		DFC	RandomForest	0.9667

Although the results of the best classifiers for each pharmaceutical compound are very similar. it is observed that in most cases better R^2 is obtained for the DFC than for the CBZ, and the best classifier is random forests. When we only consider one entry the best technique is Random Forest, in case we consider two entries, the best techniques globally are random forest and SVM. Given that the interest is to perform the least laboratory analysis possible, it is very positive that the Random Forest technique obtains satisfactory models for a single entry globally, since this model can be used to predict any entry globally by simply performing a laboratory analysis.

Table 2. Best method for each component for two inputs.

Inputs		***Output***	***Compound***	***Method***	***R²***
Concentration	Leaves	Root	*CBZ*	*SVM*	*0.9715*
			DFC	*M5Rules*	*0.9813*
Concentration	Root	Leaves	*CBZ*	*SVM*	*0.9695*
			DFC	*ANN*	*0.9705*
Concentration	Cultivars	Root	*CBZ*	*RandomForest*	*0.9822*
			DFC	*M5Rules*	*0.9812*
Concentration	Cultivars	Leaves	*CBZ*	*SVM*	*0.9836*
			DFC	*RandomForest*	*0.9784*
Leaves	Root	Concentration	*CBZ*	*RandomForest*	*0.9986*
			DFC	*RandomForest*	*0.9993*

5. Conclusions and Future Work

The problem presented in this work derives from the high environmental and economic costs involved in performing analyses to calculate the uptake of pharmaceutical compounds, in addition to the low efficiency if a large amount of analysis wants to be performed. To solve this problem in this work we have presented a study to analyze five machine learning techniques with the aim of predicting the uptake of carbamazepine and diclofenac in reclaimed water-irrigated lettuces. The study has taken into account the variables of water concentration of these two compounds, the cultivars and the absorption of these compounds both in the root and in the leaf of the lettuce. The results obtained have been very satisfactory as we have obtained robust models and a good fit of the data to the models. Better models are obtained to predict diclofenac than carbamazepine, although the difference is not significant. Therefore the overall results are optimistic and although with two input variable a little more adjustment of the models is obtained, really the difference is not significant so the models for a single input variable will be used. This means that only laboratory analyses will have to be carried out for the variable taken as input to build the model, thus achieving economic and environmental savings. The best technique for configuring the model is the ensemble Random Forest that obtains better results for carbamazepine and diclofenac on a global level.

As future work the vision carried out in this work can be applied to other types of crops. Also by means of techniques of machine learning could be analyzed not the absorption of the drug in the leaves and/or roots in reclaimed water-irrigated lettuces but also the metabolites that are detached from these compounds.

Acknowledgments

This work is supported by the Entidad Regional de Saneamiento y Depuración de Aguas Residuales (ESAMUR) [CFE-ESAMUR 37/17] and by the Spanish MINECO under grant TIN2016-78799-P (AEI/FEDER, UE) and RTC-2017-6389-5.

References

[1] Environmental Protection Agency (EPA), United States. https://www.epa.gov/laws-regulations.

[2] R.P. Bax, Antibiotic resistance: a view from the pharmaceutical industry. *Clinical Infectious Diseases (Supplement 1)* **24** (1997), S151-S153.

[3] E. Bacci, D. Calamari, C.V.M Gaggi, Biocentration of organic chemical vapors in plant leaves: experimental measurements and correlation, *Environmental Science & Technology* **24** (1990), 885-889.

[4] C. Fernández-López, J.M. Guillén-Navarro, J.J. Padilla, J.R. Parsons, Comparison of the removal efficiencies of selected pharmaceuticals in wastewater treatment plants in the region of Murcia, Spain. *Ecological Engineering* **95** (2016), 811–816.

[5] A.L. Grindlay, M. Zamorano, M.I. Rodríguez, E. Molero, M.A Urrea, Implementation of the european water framework directive: integration of hydrological and regional planning at the Segura River basin, southeast Spain, *Land Use Policy* **28** (2011). 242–256.

[6] J. Aparicio, L. Candela, O. Alfranca, J.L. García-Aróstegui, Economic evaluation of small desalination plants from brackish aquifers. Application to Campo de Cartagena (SE Spain), *Desalination* **411** (2017). 38–44.

[7] CHS, Plan Hidrológico de la Cuenca del Segura. Confederación Hidrográfica del Segura. Memoria. Ministerio de Medio Ambiente, Madrid. 2016.

[8] M. Goldstein, M. Shenker, B. Chefetz, Insights into the Uptake Processes of Wastewater-Borne Pharmaceuticals by Vegetables, *Environmental Science & Technology* **48** (2014), 5593−5600.

[9] P. A. Herklotz, P. Gurung, B. V. Heuvel, C. A. Kinney, Uptake of human pharmaceuticals by plants grown under hydroponic conditions, *Chemosphere* **78** (2010), 1416−1421.

[10] R. Tanoue, Y. Sato, M. Motoyama, S. Nakagawa, R. Shinohara, K. Nomiyama, Plant Uptake of Pharmaceutical Chemicals Detected in Recycled Organic Manure and Reclaimed Wastewater, *Journal of Agricultural and Food Chemistry* **60** (2012), 10203−10211.

[11] X. Wu, J. L. Conkle, F. Ernst, J. Gan, Treated Wastewater Irrigation: Uptake of Pharmaceutical and Personal Care Products by Common Vegetables under Field Conditions. *Environmental Science & Technology* **48** (2014), 11286−11293.

[12] F. Albani, R. Riva A. Baruzzi, Carbamazepine Clinical Pharmacology: A review, *Pharmacopsychiatry* **28** (1995), 235−244.

[13] T. Reemtsma, S. Weiss, J. Mueller, M. Petrovic, S. Gonzalez, D. Barcelo, F. Ventura, T. P Knepper, Polar Pollutants Entry into the Water Cycle by Municipal Wastewater: A European Perspective, *Environmental Science & Technology* **40** (2006), 5451−5458.

[14] T. A. Ternes, M. Bonerz, N. Herrmann, B. Teiser, H. R. Andersen. Irrigation of treated wastewater in Braunschweig, Germany: An option to remove pharmaceuticals and musk fragrances. *Chemosphere* **66** (2007), 894−904.

[15] H. Buser, T. Poiger, M.D. Mueller, Occurrence and fate of the pharmaceutical drug diclofenac in surface waters: rapid photodegradation in a lake, *Environmental Science & Technology* **32** (1998), 3449–3456.

[16] L. Lonappan, S. K. Brar, R. Kumar Das, M. Verma, R. Y. Surampalli, Diclofenac and its transformation products: Environmental occurrence and toxicity - A review, *Environment International* **96** (2016), 127–138

[17] P. A. Herklotz, , P.; Gurung , B. V Heuvel.; C. A Kinney, Uptake of human pharmaceuticals by plants grown under hydroponic conditions, *Chemosphere* **78**. (2010), 1416−1421.

[18] M Shenker, D. Harush, J. Ben-Ari, B. Chefetz,). Uptake of carbamazepine by cucumber plants—a case study related to irrigation with reclaimed wastewater, *Chemosphere* **82** (2011), 905–910.

[19] C. Riemenschneider B. Seiwert M. Moeder D. Schwarz , T. Reemtsm, Extensive Transformation of the Pharmaceutical Carbamazepine Following Uptake into Intact Tomato Plants, *Environmental Science & Technology* **51** (2017), 6100-6109.

[20] L.K. Dodgen, J., Li, D., Parker, J.J. Gan, Uptake and accumulation of four PPCP/EDCs in two leafy vegetables, *Environmental Pollution* **182** (2013), 150-156.

[21] I. Martínez-Alcalá, F. Pellicer-Martínez, C. Fernández-López, Pharmaceutical grey water footprint: Accounting, influence of wastewater treatment plants and implications of the reuse, *Water Research* **135** (2018), 278–287.

[22] M. González García, C. Fernández-López, F. Pedrero-Salcedo, J.J. Alarcón. Absorption of carbamazepine and diclofenac in hydroponically cultivated lettuces and human health risk assessment, A*gricultural Water Management* **206** (2018), 42–47.

[23] X. Wu, J. L. Conkle, J. Gan, Multi-residue determination of pharmaceutical and personal care products in vegetables, *Journal of Chromatography A* **1254** (2012b), 78–86.

[24] X. Wu, F. Ernst, J. L. Conkle, & J. Gan. Comparative uptake and translocation of pharmaceutical and personal care products (PPCPs) by common vegetables. *Environ International* , **60** (2013)15-22.

[25] Y. Wang, and I.H. Witten. Inducing model trees for continuous classes. In Proceedings of the Ninth European Conference on Machine Learning , (1997), 128-137

[26] G. Holmes, Hall, M. and Prank, E. Generating rule sets from model trees. *In Australasian Joint Conference on Artificial Intelligence*, (1999), 1–12.

[27] L. Breiman . Random Forests. *Machine learning* **45** (2001), 5-32.

[28] J..R.Quinlan. C4. 5: programs for machine learning, *Elsevier* 2014.

[29] C.E. Rasmussen and C.K. Williams. Gaussian process for machine learning. *MIT press*, 2006.

[30] D.J.C. MacKay. Introduction to gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, (1998), 133–166.

[31] Cortes, C., & Vapnik, V..Support-vector networks. *Machine learning*, **20** (1995) 273-297.

[32] Zhou, D. X., & Jetter, K.. Approximation with polynomial kernels and SVM classifiers. *Advances in Computational Mathematics*, **25** (2006) 323-344.

[33] C.M. Bishop,. Neural networks for pattern recognition. Oxford university press., 1995.

[34] A. Abraham. Artificial neural networks. Handbook of measuring system design, 2005.

[35] R. Hecht-Nielsen,. Theory of the backpropagation neural network. *In Neural networks for perception* (1992), 65-93.

APPENDIX

Table 3. Results obtained for the prediction of Carbamazepine by the different techniques taking as input one variables.

Input	*Output*	*Technique*	*r*	R^2	*STD*
Concentration	Leaves	M5Rules	0,9854	*0,9710*	0,0063
		Random Forest	0,9834	0,9670	0,0069
		SVM	0,9854	*0,9710*	0,0063
		ANN	0,9824	0,9651	0,0072
		Gaussian Process	0,9854	*0,9710*	0,0063
Concentration	Root	M5Rules	0,9865	*0,9732*	0,0078
		Random Forest	0,9866	*0,9733*	0,0077
		SVM	0,9865	*0,9732*	0,0078
		ANN	0,9851	*0,9704*	0,0082
		Gaussian Process	0,9865	0,9732	0,0078
Leaves	Concentr.	M5Rules	0,9951	*0,9902*	0,0026
		Random Forest	0,9993	*0,9986*	0,0012
		SVM	0,9854	0,9710	0,0063
		ANN	0,9817	0,9637	0,0093
		Gaussian Process	0,9854	0,9710	0,0063
Leaves	Root	M5Rules	0,9764	0,9534	0,0107
		Random Forest	0,9825	0,9653	0,0092
		SVM	0,9764	0,9533	0,0106
		ANN	0,9729	0,9466	0,0090
		Gaussian Process	0,9764	0,9533	0,0106
Root	Concentr	M5Rules	0,9963	0,9926	0,0018
		Random Forest	0,9999	0,9999	0,0003
		SVM	0,9865	0,9732	0,0078
		ANN	0,9893	0,9787	0,0054
		Gaussian Process	0,9865	0,9732	0,0078

Root	Concentr	M5Rules	0,9963	0,9926	0,0018
		Random Forest	0,9999	0,9999	0,0003
		SVM	0,9865	0,9732	0,0078
		ANN	0,9893	0,9787	0,0054
		Gaussian Process	0,9865	0,9732	0,0078
Root	Leaves	M5Rules	0,9746	0,9499	0,0099
		Random Forest	0,9809	0,9622	0,0063
		SVM	0,9764	0,9533	0,0106
		ANN	0,9796	0,9597	0,0058
		Gaussian Process	0,9764	0,9533	0,0106

Table 4. Results obtained for the prediction of Carbamazepine by the different techniques taking as input two variables.

Input		Output	Technique	r	R^2	STD
Concentr	Leaves	Root	M5Rules	0.9852	0.9706	0.0081
			Random Forest	0.9826	0.9656	0.0093
			SVM	0.9857	0.9715	0.0082
			ANN	0.9843	0.9688	0.0083
			Gaussian Process	0.9868	0.9739	0.0077
Concentr	Root	Leaves	M5Rules	0.9820	0.9643	0.0081
			Random Forest	0.9809	0.9622	0.0063
			SVM	0.9846	0.9695	0.0051
			ANN	0.9809	0.9621	0.0052
			Gaussian Process	0.9856	0.9714	0.0054
Concentr	Cultiv.	Root	M5Rules	0.9845	0.9693	0.0076
			Random Forest	0.9911	0.9822	0.0064
			SVM	0.9850	0.9703	0.0075
			ANN	0.9890	0.9780	0.0060
			Gaussian Process	0.9807	0.9618	0.0114
Concentr	Cultiv.	Leaves	M5Rules	0.9913	0.9826	0.0041
			Random Forest	0.9892	0.9785	0.0062
			SVM	0.9917	0.9836	0.0041
			ANN	0.9893	0.9788	0.0041
			Gaussian Process	0.9862	0.9725	0.0075
Leaves	Root	Concentr	M5Rules	0.9951	0.9902	0.0026
			Random Forest	0.9993	0.9986	0.0012
			SVM	0.9913	0.9827	0.0049
			ANN	0.9890	0.9781	0.0057
			Gaussian Process	0.9914	0.9830	0.0053

Table 5 Results obtained for the prediction of Diclofenac by the different techniques taking as input a single variable.

Input	Output	Technique	r	R^2	STD
Concentration	Leaves	M5Rules	0.9838	0.9679	0.0074
		Random Forest	0.9866	0.9735	0.0066
		SVM	0.9838	0.9679	0.0074
		ANN	0.9853	0.9708	0.0070
		Gaussian Process	0.9838	0.9679	0.0074
Concentration	Root	M5Rules	0.9907	0.9816	0.0040
		Random Forest	0.9891	0.9784	0.0063
		SVM	0.9907	0.9816	0.0040
		ANN	0.9899	0.9799	0.0041
		Gaussian Process	0.9907	0.9816	0.0040
Leaves	Concentration	M5Rules	0.9957	0.9914	0.0027
		Random Forest	0.9996	0.9993	0.0010
		SVM	0.9838	0.9679	0.0074
		ANN	0.9838	0.9679	0.0095
		Gaussian Process	0.9838	0.9679	0.0074
Leaves	Root	M5Rules	0.9722	0.9452	0.0116
		Random Forest	0.9871	0.9744	0.0061
		SVM	0.9715	0.9438	0.0110
		ANN	0.9705	0.9418	0.0112
		Gaussian Process	0.9715	0.9438	0.0110

		M5Rules	0.9966	0.9931	0.0018
		Random Forest	1.0000	1.0000	0.0001
Root	Concentration	SVM	0.9907	0.9816	0.0040
		ANN	0.9906	0.9813	0.0064
		Gaussian Process	0.9907	0.9816	0.0040
		M5Rules	0.9723	0.9454	0.0116
		Random Forest	0.9832	0.9667	0.0075
Root	Leaves	SVM	0.9715	0.9438	0.0110
		ANN	0.9732	0.9471	0.0124
		Gaussian Process	0.9715	0.9438	0.0110

Table 6. . Results obtained for the prediction of Diclofenac by the different techniques taking as input two variables.

Input		**Output**	**Technique**	**r**	**R^2**	**STD**
Concentr	Leaves	Root	M5Rules	0.9906	0.9813	0.0042
			Random Forest	0.9872	0.9745	0.0071
			SVM	0.9888	0.9778	0.0043
			ANN	0.9891	0.9782	0.0043
			Gaussian Process	0.9891	0.9782	0.0046
Concentr	Root	Leaves	M5Rules	0.9835	0.9673	0.0075
			Random Forest	0.9833	0.9668	0.0075
			SVM	0.9824	0.9652	0.0077
			ANN	0.9852	0.9705	0.0067
			Gaussian Process	0.9816	0.9634	0.0079
Concentr	Cultivars	Root	M5Rules	0.9905	0.9812	0.0037
			Random Forest	0.9853	0.9709	0.0072
			SVM	0.9903	0.9806	0.0037
			ANN	0.9805	0.9613	0.0095
			Gaussian Process	0.9848	0.9699	0.0077
Concentr	Cultivars	Leaves	M5Rules	0.9829	0.9660	0.0082
			Random Forest	0.9891	0.9784	0.0048
			SVM	0.9820	0.9643	0.0069
			ANN	0.9849	0.9700	0.0068
			Gaussian Process	0.9775	0.9556	0.0161
Leaves	Root	Concentr	M5Rules	0.9957	0.9914	0.0027
			Random Forest	0.9996	0.9993	0.0010
			SVM	0.9944	0.9889	0.0028
			ANN	0.9940	0.9881	0.0041
			Gaussian Process	0.9946	0.9893	0.0026

Agriculture and Environment Perspectives in Intelligent Systems
A. Muñoz and J. Park (Eds.)
IOS Press, 2019

doi:10.3233/AISE190006

Minimum Temperature Prediction Models in Plots to Forecast Frost in Crops

M. Ángel GUILLÉN-NAVARRO [a,1], Jose M. CADENAS [b], M. Carmen GARRIDO [b]
Belén AYUSO [a] and Raquel MARTÍNEZ-ESPAÑA [a]

[a] *Dept. of Computer Engineering, Catholic University of Murcia, Murcia, Spain*
[b] *Dept. of Information and Communications Engineering, University of Murcia, Murcia, Spain*

Abstract. The concept of precision agriculture tries to integrate the problems of agriculture with new technologies, in order to provide effective, feasible and efficient solutions, always trying to obtain a higher yield from farmers. Agriculture is an area of high economic relevance in many places. Specifically in this study we will focus on the Spanish province of the Region of Murcia, where agriculture accounts for more than 20% of its economy. Frosts in crops are posing as a serious problem for farmers in this area due to climate change. In this paper we address the problem of frosts suffered by farmers in southeastern Spain with frosts in their crops. At the end of the winter season, temperatures vary by as much as 20 degrees Celsius from midday to night. These variations provoke the anticipation of the blossoming in stone fruit trees, having the risk of frosts at night.Thus, in this paper, the Intelligent Data Analysis have been applied to create predictive models of minimum temperature in plots. In addition, a selection process of the most relevant characteristics to predict the minimum temperature will be presented using the information provided by the models. The data necessary to carry out this study will be collected from the different weather stations of the Institute of Agricultural and Food Research and Development of Murcia. Specifically, data from forty weather stations have been studied, with the aim of finding local or global models that predict the temperature one hour in advance. The data analysis techniques used for the prediction models have been the M5 rule technique for predicting the minimum temperature and the C4.5 decision tree for classifying whether frost will occur or not. The results have identified the most relevant attributes both for predicting and characterize temperature and for classifying whether frost: dew point, vapour pressure deficit and maximum relative humidity occur. The results obtained indicate that both a local classification model and a local prediction model fit perfectly to the resolution of the problem obtaining on the one hand an error of less than 0.5 degrees Celsius for the prediction of the minimum temperature and on the other hand a precision of 98% for the classification of whether frost will occur or not.

Keywords. precision agriculture, crop frost, data mining, intelligent data analysis

[1] Corresponding Author: M. Ángel, Guillén-Navarro, Dept. of Computer Engineering, Catholic University of Murcia, Murcia, Spain; E-mail:maguillen@ucam.edu

1. Introduction

In agriculture, the use of new technologies is being implemented to improve productivity, efficiency and cost savings. These new technologies offer a collection of information of various types (graphics, sensors, positioning, etc.) in a fast way that allows to take decisions and act at the right time. Until recently, capturing information from a farmer usually consisted of seeing how the crop grew in the neighbor's field. Agriculture was based on observing the behavior of the fields based on the experience of older farmers. Farmers also relied on the advice of agronomists, but no analyses were carried out on the type of soil, type of crop or farming area to solve problems more precisely, more efficiently and more environmentally responsible. Currently, on the medium-sized farm, the system has not changed much, and although it is an information capture and management system that can function fairly well, it is far from effective. In the midst of the digital age agriculture area is one step behind in these computational issues, being the availability and management of information at least as important as in other sectors such as industry or services [4]. Spanish farmers were generally very reluctant to take systematic data on their farm, as they understood this process to be costly in time and resources with no short-term reward, so they considered it as an uninteresting investment. However, more and more farmers are interested in modernizing their crops using new technologies. For this they are allowing the installation and implementation of applications of sensors, cameras, drones, among other elements that utilize the communications and advantages offered by Internet of Things [19]. In addition to the advantages offered by the Internet of Things, there are good practices and techniques offered by the discipline of precision agriculture. Precision agriculture uses information and communication technology together with best agricultural practices for farm management [29]. This discipline requires a process of acquiring, transmitting and processing large amounts of field data. In order to be able to understand the data acquired from the farm, it is necessary to analyze and study these data in order to create models that can help farmers in their decision making. Intelligent Data Analysis (IDA) is a discipline composed of several phases that allows manipulating, studying and preparing data to obtain intelligible models that are capable of being the basis of decision making to solve problems (these phases can be summarized in: a) "Data collection", b) "Data preprocessing" - encompassing the data preparation, data transformation, etc-, c) "Data Mining" and d) "Evaluation" - referring to the evaluation and interpretation of models-). Therefore, this discipline will be used in this study to confront the problem of frost in crops.

In southeastern Spain, farmers have recently suffered heavy crop losses due to inclement weather. Specifically, the damage is occurring due to sudden changes in temperature, during the daytime temperatures exceed 15 degrees Celsius and at night negative temperatures occur. These temperature changes cause stone fruit trees to bloom prematurely and the negative temperatures end up by freezing these flowers, resulting in fruit loss [5,6]. Specifically, this paper is focused on stone fruit trees. To mitigate the effects of a frost in stone fruit trees, there are some techniques that can reduce losses depending on the intensity of the frost. Among techniques to mitigate these effects are included the following:

- Protective covers: The covers can be plants, straw or plastics.
- Artificial fog: This fog is produced by forming smoke by burning cheap waste (straw and by-products) or chemicals such as paraffin.

- Fans: This technique is based on air agitation to break the thermal inversion. The use of fans mixes hot and cold air and breaks the thermal inversion.
- Stoves: The cold protection action is due to two complementary phenomena: the emission of infrared radiation when the device is hot and the heating of the air by conduction.
- Sprinkler irrigation: This technique is applied by watering the crop with sprinklers before and during frost because the water freezes the heat and the temperature does not drop below 0 Celsius degrees, but this must be continuous throughout the event.
- Application of chemical treatments: This technique is applied to cause delays in the development of crops or to increase the cold resistance of plants/flowers.

The problem that arises with crop frost is the need to know in advance if a frost will occur or not to prepare the material or workers needed to combat frost using any of the techniques described. In this paper several IDA techniques are applied to help us to predict frosts in stone fruit trees. For this purpose, data from different weather stations are used in order to predict, with the least possible error, when a frost will occur. In addition, an analysis of the models is performed to determine the most important climatic variables to predict the temperature. The techniques applied help us to define elements that allow us to build a decision support system for farmer.

This paper is organized as follows: Section 2 presents a review of the works related to precision agriculture and how the IDA process is applied in this field. Section 3 describes the data used for temperature prediction to detect frost and the IDA techniques used to make that prediction. Furthermore in this section results are also shown and discussed. Finally in Section 4 conclusions and future work are presented.

2. Background

Precision agriculture is a concept that was born in the mid-1980s [15], although it is being used in recent years to a greater extent. This concept involves the use of technologies to help solve agricultural problems. Precision agriculture allows us to perform agricultural tasks correctly, in the right place, at the right time and in the right way [16]. Precision agriculture generally involves a better management of agricultural inputs such as fertilizers, herbicides, seed, fuel, etc. This type of agriculture, unlike conventional agriculture, offers a personalized attention to each type of crop. In this way, precision agriculture trend to optimize the profitability, sustainability and to reduce the environmental impact maximizing the crop productivity [13]. Thus, precision agriculture can be defined as a production system that promotes variable management practices within a field, according to site conditions. This system uses tools and information sources that can be provided by various technologies such as global positioning system, geographic information systems, yield monitoring devices, soil, plant and pest sensors, remote sensing, and variable-rate technologies [22]. In addition to the technologies used for data collection, it is important to consider the techniques needed to analyze all the information collected to provide rules, actions or decisions for better results in the precision agriculture process. Data mining (DM) offers a set of computational techniques to treat and manage these amounts of data and to solve the specific and general problems that appear in agriculture [14]. More and more studies are using DM techniques to solve problems in agriculture such as species recognition, disease detection, yield prediction, disease detection, weed detection or crop quality [30].

In [17] a perception system for agricultural robotics with a multispectral camera is presented to automatically perform the tasks of detection and classification of crops and weeds in real time. Two different convolution neural networks are applied. The first for segmentation of camera images and the second to classify between crops and weeds using the pixels extracted from that segmentation. In this study authors present the components needed to capture the data as well as the DM techniques to analyze them. In [24] a visible and near infrared reflectance spectroscopy is used to collect data on soil organic carbon. These data are pre-processed using the random forest technique and analyzed by multivariate regression models in order to predict soil organic carbon. In [32] the authors propose a technique to solve the problem of crop selection and maximize the rate of net yield in the crop throughout the season. Another work that uses DM techniques for performance prediction is presented in[33]. In this study a deep neural network is applied to estimate corn yield. In addition, an evaluation of the prediction of the estimation of the obtained model is carried out comparing it with a support vector machine (SVM). The authors of [34] have considered climate effects, soil salinity and production area as key factors for crop production in Bangladesh. Taking these factors into account, they have first used clustering techniques to group regions and then used DM techniques such as K-nearest neighbors and neural networks to obtain crop yield predictions.

Predicting diseases in crops is another area where precision agriculture can be used. Thus, the authors in [25] propose the development of a model to predict groundnut pest, specifically Thrips and Bud Necrosis diseases. To achieve this goal, the authors deployed a network of sensors to obtain real-time weather parameters such as temperature, humidity and leaf wetness. Using these data, the authors propose the use of DM techniques such as Gaussian Naive Bayes and Rapid Association Rule Mining to detect hidden correlations in data that help detect these diseases in crops. Another paper where DM techniques are applied to detect plant diseases is presented in [2]. In this study an automatic detection and classification of leaf diseases has been proposed. This method is based on K-means as a clustering procedure and a neural network as a sorting tool using a set of texture characteristics.

DM techniques have also been applied to weather forecasting from data obtained from meteorological stations. Within this scope there are works in which the prediction of temperature, wind speed, abnormal events such as hurricanes or storms, prediction of fog or forecast of rainfall is carried out. This weather prediction is important when making decisions in agriculture, such as estimating the probability of fires, crop diseases or managing elements such as irrigation. Thus, among the other works that carry out the temperature prediction, in [12] authors perform the estimation of the average, minimum and maximum temperature of one day from the data of previous years. They carry out regression through k-nearest neighbor (kNN), decision trees, rules, neural networks and additive regression. In [20] from the maximum temperature of several days earlier, predict the next day's temperature. Authors take the maximum daily temperature of a dataset by performing the imputation of unknown values. The prediction is made with a SVM and a multilayer perceptron obtaining better results with SVM. In [10] the kNN technique is used to predict the weather (temperature, humidity, rain, ...) of a region from the nearest neighbors in a historical dataset. Rainfall plays a very important role in agriculture, especially in areas where water is a scarce resource and good management is necessary, hence the importance of predicting it and therefore there are works focused on this objective. Thus in [23] a multiple linear regression technique for the early prediction of rainfall is developed. In [11] the authors carry out the prediction using a Bayesian classifier obtaining good results with 7 attributes for large datasets. The authors of [31] compare the model of Markov chain extended with rainfall prediction with six DM techniques (Genetic Programming, M5 Model trees, Radial Basis Neural Networks, M5 Rules, Support Vector Regression, and k-Nearest Neighbours). To carry out this study they use on the one hand the daily rain prediction and on the other hand the accumulated precipitation each certain time window.

Short-term weather warnings are becoming increasingly important. Predicting the existence of low cloud cover is important in agriculture as a local flood warning. To this end, the authors in [3] develop an study of fog and low cloud cover prediction from meteorological data using CRISP-DM methodology. In the paper they only carry out the initial stages of the methodology, such as data analysis and cleaning. Although the reliability of weather stations has improved recently, there are still faults that lead to data loss. Thus, in [1] the imputation of these values using data obtained from neighbouring stations is performed.

3. Intelligent data analysis process for predicting frost crop in agriculture

This section describes the different phases of the IDA process that have been carried out with the aim of studying various useful elements for the implementation of a decision support system for the farmer. This system should help farmers to take the right actions to protect their crops from frost. The IDA process starts from the meteorological dataset available from several weather stations in a close environment and after performing a data preprocessing to obtain the minable view. Two DM techniques are applied, one based on decision tree and other based on decision rules. One of them solves the regression task and another solves the classification task. The objective of the classification task is indicate to the farmer if there is going to be a frost or not, while the regression task will indicate to the farmer the minimum temperature that is expected within an hour. After constructing the classification and regression models, the results obtained are analized and the most influential variables in the construction of both models are studied.

3.1. Data Collecting and Preprocessing

The dataset used for this study have been obtained from forty meteorological stations belonging to the Institute of Agricultural and Food Research and Development of Murcia, [9]. Although this organization has more stations these stations have been selected because they are the most equipped and informative they have. In addition these stations are located in various areas of the region of Murcia which allows us to study the difference and variability of conditions between areas. The location and acronym used for each of these stations is described in Table 5. Each station is equipped with the following sensors and ephemeris: weather vane, radiometer, rain gauge, data-logger and thermo-hygrometer.

3.1.1. Data used

The dataset used for this study range from 1 January 2012 to 30 April 2018 for each weather station. For each day there are a total of 24 values, one for each hour. Each instance of the dataset, used for this study, is composed of 16 attributes detailed in Table 1. As it can be seen in this table, the meteorological attributes used range from temperature and humidity to radiation, including precipitation and variables related to wind.

Table 1. Attribute Description

Attribute	Description	Attribute	Description
STATCODE	Weather station code	MEANWS	Mean wind speed
DATE	Date of data reading	MEANWD	Mean wind direction
HOUR	Hour of data reading	MAXWS	Maximum wind speed
RHMEAN	Mean relative humidity	RFALL	Rainfall
RHMAX	Maximum relative humidity	DEWPT	Dew point
RHMIN	Minimum relative humidity	VPD	Vapor pressure deficit
MEANRAD	Mean radiation	TMIN	Minimum temperature
MAXRAD	Maximum radiation	RADACU	Accumulated radiation

This dataset containing the information of all the stations is named as DS_{all}. The size of this dataset is stipulated in Table 2. In addition to studying all the information as a whole model, the forty stations have been grouped to make a total of nine groups. The stations have been grouped by their proximity. The groups will be indicated in the station code, since stations whose code begins with the same letters are grouped. Therefore, each individual dataset is going to be labeled as DS_{XX}, with XX being the acronym for group to belong it.

Figure 1 depicts the map of the Region of Murcia (Spanish province) where the grouped stations are visualized with the same letters in the code of the station. As it can see the stations are grouped around different populations. It should be borne in mind that there are two stations that have not considered due to the scarcity of data they present because they do not have a complex equipment to measure weather variables.

3.1.2. Data Preprocessing

Given that the main objective of the IDA process is the frost crop prediction, first a selection of instances is carried out to obtain a data subset describing situations close to frost. In this way, all instances with a value TMIN > 7.0 are eliminated. This filter is carried out to avoid having temperatures that are irrelevant to the domain of the problem to study. Also, there are some attributes that are irrelevant for the analyzed problem, such as the attributes STATCODE, DATE and HOUR. These attributes have been eliminated. In addition, the attribute RADACU have also been eliminated due to in more than 50% of the stations this attribute had no values and did not provide relevant information.

Regarding the instances, those instances that did not have information in at least eight of the attributes have been eliminated since these instances did not provide information and were a source of noise to obtain a quality model.With respect to the instances with less than eight missing values, they have been conserved since the DM techniques that we are going to apply allow the manipulation of this type of values.

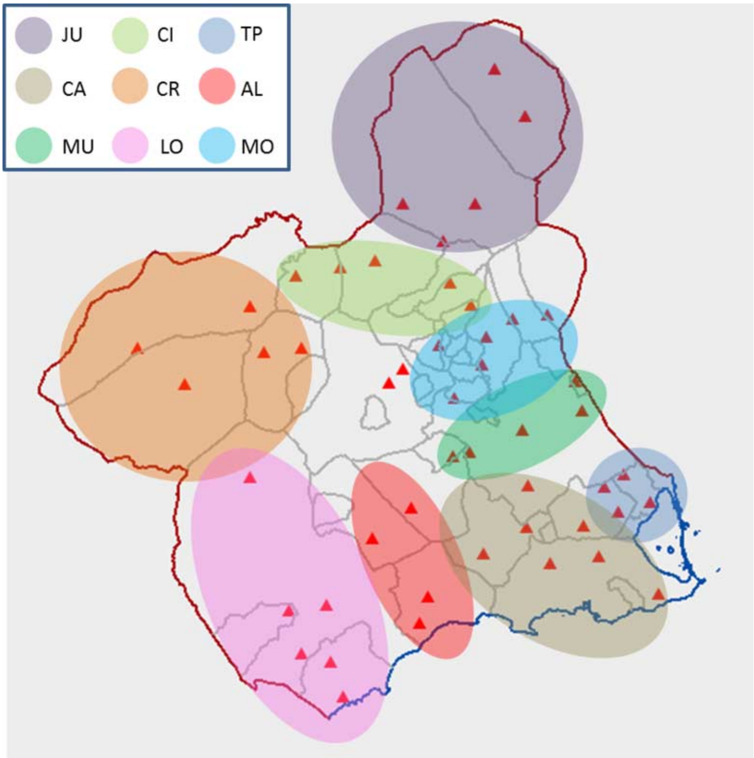


Figure 1. Map of the region of Murcia (Spain) showing the grouping of the weather stations.

On the other hand, the datasets considered in this work contain information that is stored chronologically in constant time periods. They contain information with special characteristics due to the temporal relationships between the data. These characteristics can not be directly addressed by traditional DM techniques [8]. But, there are two ways to approach this problem:

- The development of specific techniques to address this kind of information without any transformation.
- The transformation of information so that it reflects these intrinsic characteristics and can be treated by non-specific DM techniques.

In this work the second option is followed, carrying out a data transformation that captures this temporal relationship using traditional DM techniques. Thus, the initial attribute set is extended with one numerical attribute and one nominal attribute:

- TMIN_{DIF}: which indicates the difference between the minimum temperatures of the two readings prior to the current one.
- CLASS: label assigned to the instance that can take two possible values $\{$FROST, NOFROST$\}$. To carry out the label assignment, the following decision rule has been applied: "if TMIN$\leq$ 0 then CLASS=FROST else CLASS=NOFROST".

Before going on to the next phase of the IDA, the minable views of the nine groups of stations are constructed (DS_{AL}, DS_{CA}, DS_{CI}, DS_{CR}, DS_{JU}, DS_{LO}, DS_{MO}, DS_{MU} and

DS_{TP}) to later form the complete dataset (DS_{all}) with all the information of all the stations studied. The dataset description that constitutes the minable view of data is shown in the Table 2.

Table 2. Dataset Description

		Input Attributes	
Dataset	Instances	Numerical	Nominal
DS_{all}	199235	14	1
DS_{AL}	16842	14	1
DS_{CA}	18585	14	1
DS_{CI}	18577	14	1
DS_{CR}	36241	14	1
DS_{JU}	36310	14	1
DS_{LO}	22691	14	1
DS_{MO}	15338	14	1
DS_{MU}	14229	14	1
DS_{TP}	20422	14	1

3.2. Data Mining and Evaluation

Once the data minable view is obtained, the DM phase applies techniques on these data in order to solve regression or classification tasks. In this paper, two techniques are applied to the available data: a technique of construction of decision rules to carry out the estimation of the numerical attribute TMIN (regression task) and a classification tree technique to carry out the estimation of the nominal (categorical) CLASS attribute (classification task). The implementations of these techniques (M5Rules and J48 for the regression and classification, respectively) are provided by Weka package [7]. Decision trees and decsion rules model the linear/nonlinear relationships between the attributes and, in addition, they are interpretable and understandable.

In general, techniques generate decision trees for a dataset by recursive partitioning. A decision tree is grown using a depth-first strategy. Techniques consider all the possible attributes that can split the data in each node and select the one that gives the best value for a given measurement. The process continues until the nodes verify a stop condition in which case these nodes become leaf nodes representing a decision. From a built decision tree, the inference process for a new instance starts at the tree root. The instance is then evaluated at one node and takes the appropriate branch to its outcome. It continues through several internal nodes until it finds a final leaf. The decision is represented by these leaves nodes. J48 implementation is a C4.5 decision tree. To build each node, C4.5 [18] selects the attribute that gives the best information gain. Leaves represent the values (labels or categories) of the class attribute.

Different techniques exists to build the regression model. Some techniques utilize separate-and-conquer rule learning and the more popular techniques learn regression trees which can then be converted into decision rules. Between the latter, M5Rules [27] implementation is found. The key idea of these latest techniques is to replace the purity heuristic of the decision tree algorithm with a heuristic that measures the reduction in variance. M5Rules technique uses the decision tree M5P [26]. Leaves of the M5P trees are composed of multivariate linear regression models and nodes are built over the attributes that minimize the expected error (deviation) respecting the output attribute. In each iteration of the M5Rules technique, a decision tree is built using M5P technique and the "best" leaf is turned (the one with the highest coverage) into a rule and the tree used is discarted. Thus, M5Rules generates a list of regression decisions using separate-and-conquer strategy or covering strategy.

To carry out the inference of the TMIN and the CLASS attributes different subsets of the datasets described above have been considered. The subsets are selected in order to analyze how important is the temporal relationship in the data and the importance of the mean values against the extreme values of a measure. With this information we can understand better the intrinsic relationships in the data and improve the data collection phase, focusing it on the collection of the most relevant values for the problem in study. The subsets are obtained as follows (in all of them the CLASS attribute is eliminated when the regression task is carried out and the TMIN attribute is eliminated when the classification task is carried out):

Conf$_1$: Configuration with the highlighted attributes of the Table 1 by adding the TMIN/CLASS and TMIN$_{\mathrm{DIF}}$ attributes (14 attributes).

Conf$_2$: Configuration with the attributes of Conf$_1$ where MEANRH, MEANRAD and MEANWS attributes are eliminated (so that only the extreme values of these measures are used in the instances) (11 attributes).

Conf$_3$: Configuration with the attributes of Conf$_1$ where MEANRH, MEANRAD, MEANWS and TMIN$_{\mathrm{DIF}}$ attributes are eliminated (10 attributes).

3.2.1. Regression task

To carry out the inference of the TMIN attribute, the M5Rule technique has been applied to the different configurations of the ten datasets. The measures used to analyze the results are the following:

- the root mean square error (RMSE)
- the mean absolute error (MAE)
- the Pearson correlation coefficient (CC) that measures the statistical correlation between predicted data (Y) and real ones (X). It takes values between [-1,1]. CC is defined according to $CC(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$ where $Cov(X,Y)$ is the covariance between X and Y, and σ_X, σ_Y are the standard deviations of X and Y, respectively. The different possibilities from this parameter are the following:
  - If CC= 1: total dependency between predicted and real values (direct correlation).
  - If 0 <CC< 1: positive correlation.
  - If CC= 0: not lineal correlation.
  - If -1 <CC< 0: negative correlation.
  - If CC= -1: total dependency between both values (inverse correlation).
- the determination coefficient (R^2) that measures how well data fit a statistical model (proportion of the variance in the dependent attribute that is predictable from the independent attributes). It is defined as $R^2 = \left(1 - \left(\frac{\sum_i (X_i - Y_i)^2}{\sum_i (X_i - \bar{X})^2}\right)\right)\%$.

The results obtained are shown in the Table 3. The values obtained to these measures for a 5-fold cross-validation are indicated to each dataset.

Table 3. Results with M5Rules technique and a 5-fold cross-validation

	Dataset	CC	R^2	MAE	RMSE
with Conf$_1$ (14 attributes)	DS$_{all}$	0.9887	97.75	0.2400	0.3676
	DS$_{AL}$	0.9829	96.61	0.2968	0.4108
	DS$_{CA}$	0.9851	97.03	0.2022	0.2860
	DS$_{CI}$	0.9892	97.86	0.2185	0.3069
	DS$_{CR}$	0.9937	98.75	0.2378	0.3397
	DS$_{JU}$	0.9929	98.58	0.2283	0.3228
	DS$_{LO}$	0.9749	95.04	0.2864	0.4597
	DS$_{MO}$	0.9892	97.86	0.2011	0.2919
	DS$_{MU}$	0.9595	92.06	0.3033	0.5139
	DS$_{TP}$	0.9702	94.13	0.2451	0.4045
with Conf$_2$ (11 attributes)	DS$_{all}$	0.9885	97.70	0.2416	0.3711
	DS$_{AL}$	0.9819	96.41	0.3051	0.4231
	DS$_{CA}$	0.9849	97.00	0.2039	0.2873
	DS$_{CI}$	0.9889	97.79	0.2229	0.3121
	DS$_{CR}$	0.9935	98.70	0.2415	0.3469
	DS$_{JU}$	0.9927	98.55	0.2295	0.3269
	DS$_{LO}$	0.9744	94.95	0.2866	0.4647
	DS$_{MO}$	0.988	97.61	0.2102	0.3086
	DS$_{MU}$	0.9591	91.99	0.3082	0.5162
	DS$_{TP}$	0.9683	93.76	0.2534	0.4173
with Conf$_3$ (10 attributes)	DS$_{all}$	0.9884	97.69	0.2443	0.3725
	DS$_{AL}$	0.9819	96.41	0.3048	0.4224
	DS$_{CA}$	0.9842	96.86	0.2096	0.294
	DS$_{CI}$	0.9886	97.73	0.2248	0.3162
	DS$_{CR}$	0.9934	98.68	0.2439	0.3489
	DS$_{JU}$	0.9927	98.55	0.2314	0.3265
	DS$_{LO}$	0.9737	94.81	0.2932	0.4705
	DS$_{MO}$	0.9881	97.63	0.2115	0.3068
	DS$_{MU}$	0.9602	92.20	0.3007	0.5092
	DS$_{TP}$	0.969	93.90	0.2526	0.4126

When analyzing the results it is observed that both the coefficients of correlation and determination are very good in all the models (for each model and each configuration of the dataset, the correlation between the TMIN attribute and the rest of the attributes, and the proportion of explained variance are very high).

Appreciably better models (they get higher correlation/determination coefficients with a lower error) are obtained using the DS$_{\mathrm{JU}}$ and DS$_{\mathrm{CR}}$ datasets, regardless of the configurations. This aspect is graphically visible in Figure 2, where although all the models have a great fit, the stations of DS$_{JU}$ and D$_{SCR}$ get a fit of a point of average over the rest of the models. In contrast, also graphically, in Figure 2 it can be observed that the model obtained for the group of stations DS$_{MU}$ that obtains the worst model fit followed by the groups of stations DS$_{TP}$ and DS$_{LO}$.

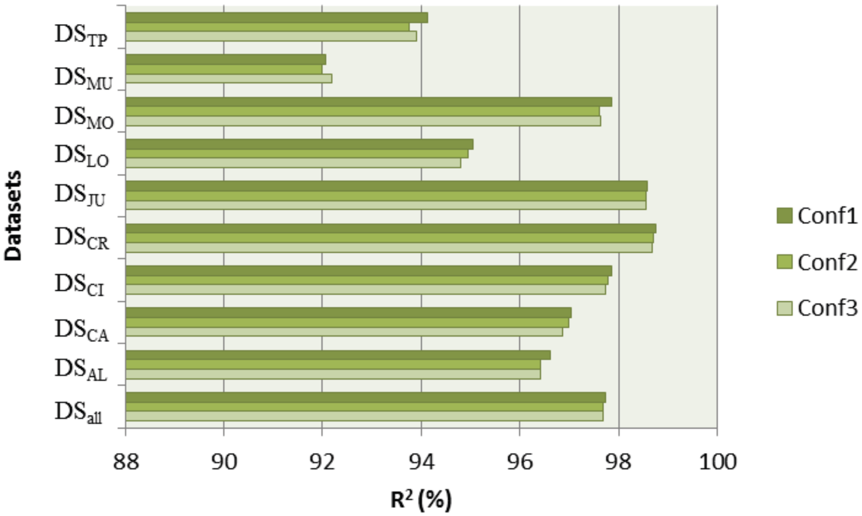


Figure 2. Comparison of the value of the coefficient of determination (R^2) for the three configurations to predict the TMIN attribute

Until now the fit of the models obtained to the data has been analized, however a fundamental aspect for the sutudied problem is the error that the model makes when predicting the minimum temperature (TMIN). Figure 3 shows a comparison of the root mean square error of the three configurations for all the datasets studied. In Figure 3 it can be seen how the smallest error is always obtained by the group of stations that DS$_{CA}$ dataset represents. However the DS$_{CR}$, DS$_{JU}$ stations, as for the R^2 measure, obtain a low error of less than 0.35 for the three configurations. The biggest error, as with R^2, is obtained by the set of stations represented by the dataset DS$_{MU}$ followed by the dataset DS$_{LO}$. The dataset DS$_{MU}$ has an error greater than 0.5 degrees Celsius which can already be a problem when making decisions because in addition to the error of the prediction technique must be taken into account the error that occurs in the measuring instruments.

Since the M5Rules technique obtains the rules by means of M5P decision trees and M5P obtains a tree model whose leaves are multivariate linear models, an analysis is carried out to know which attributes are part of the test nodes of the tree (internal nodes) and which attributes are part of the multivariate linear models of the leaves. This analysis is performed in depth on the DS$_{\text{all}}$ dataset with the three configurations defined. The analysis indicates that the models obtained with the three configurations have a similar structure and certain attributes in common. Specifically, they share the DEWPT, VPD, MAXRH and MAXRAD attributes that are used in the main tree branches. It should be noted that the DEWPT and VPD attributes appear in the main tree nodes,

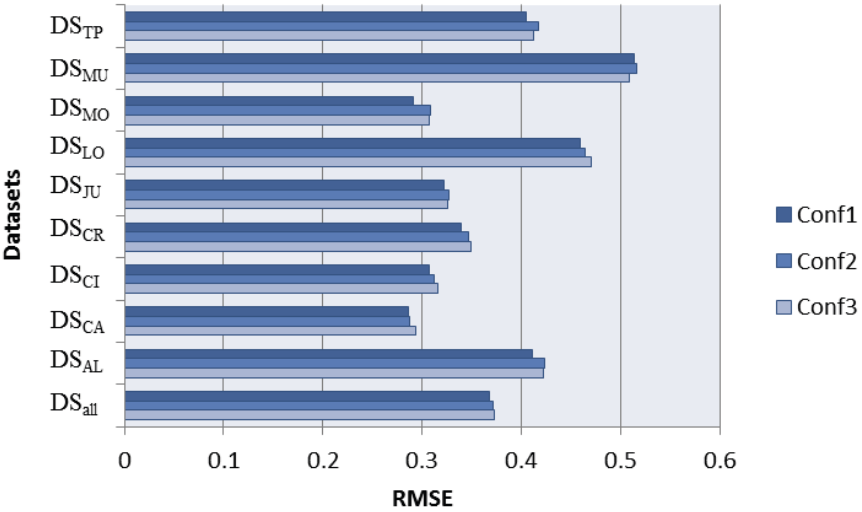


Figure 3. Comparison of the value of Root Mean Square Error (RMSE) for the three configurations to predict the TMIN attribute

indicating that they have a greater power of data discrimination for this regression task, that is, these attributes are influential in predicting the minimum temperature. Besides, it is also noteworthy that for configurations two and three the RAINFALL attribute and MEANSWS also appear as highlighted. Because these attributes are important in the models, they can also be considered important in the problem studied.

As a conclusion to this regression task can be mentioned that both local and global models can be created to predict the minimum temperature since both the model fit and the average error obtained is quite satisfactory. It is worth noting the result of the DS$_{MU}$ dataset that obtains a worse model fit and a worse error for the three configurations designed, but it is not a problem since for this group of stations the global DS$_{all}$ dataset model can be used that obtains an acceptable prediction error and a good fit.

3.2.2. Classification task

Given the results obtained in the regression task and since the global model obtains good results, in this section the inference of the attribute CLASS is carried out, applying the decision tree J48 to the DS$_{\mathrm{all}}$ dataset with the three designed configurations. To solve the classification task, the TMIN attribute is eliminated and now the CLASS attribute is considered.

The measures used to analyze the results are the following:

- The accuracy of the model.
- Area under the ROC curve of the model (the perfect model will get a value of 1).
- The f-measure (harmonic mean of precision and recall), evaluated in [0,1] (the best model will be indicated with a value of 1).

The obtained results are shown in Table 4. The table shows the average of these measures to a 5-fold cross-validation.

Table 4. Results for DS_{all} dataset with J48 technique and a 5-fold cross-validation

Configuration	Attributes	Accuracy	ROC area	f-measure
$Conf_1$	13	98.87%	0.982	0.989
$Conf_2$	10	98.88%	0.983	0.989
$Conf_3$	9	98.89%	0.986	0.989

As can be seen in the results obtained, the three models learned have a similar behavior. Although the best results are obtained by $Conf_3$, really the difference is negligible and it is possible to use any configuration to carry out a classification of whether there is frost or not.

Again (as with the regression models), analyzing the decision tree nodes, the attributes that are common to the three models are: DEWPT, VPD, MAXRH, MAXRAD, MEANWD and MAXWS. As in the regression analysis, the DEWPT and VPD attributes appear in the main nodes of the trees, indicating that they have a greater power of data discriminating for the classification task.

Considering the attributes highlighted in the estimation of the TMIN attribute and those highlighted in the classification of the CLASS attribute, it can be concluded that the attributes DEWPT, VPD, MAXRH, MAXRAD should be considered of special relevance in the problem of crop frost prediction. This shows that if a farmer wants to have a local model of his crops, he must install the necessary measuring instruments to measure these variables as accurately as possible.

The prediction results obtained in both classification and regression are very interesting in order to make a temperature prediction to prevent frost. This study indicates the most relevant climatic variables to carry out a local study in other plots and thus create a web application that alerts farmers when a frost is going to occur, either by providing the temperature or by indicating whether a frost can occur or not.

4. Conclusions and Future Work

In this study the problem of frost prediction in the region of Murcia (Spain) is faced using Intelligent Data Analysis. Data from a total of forty stations belonging to the Institute of Agricultural and Food Research and Development of Región have been collected. These data have been preprocessed to eliminate those data irrelevant to the proposed study and to add the necessary information to represent the temporality of the data, as well as to decide when a frost occurs and when it does not. The data mining techniques of the C4.5 decision tree and the M5Rules rule have been used to build the models. The results obtained are very satisfactory since an error of less than 0.5 degrees Celsius has been obtained for forty of the stations and for the global model that covers all stations.

Therefore, the models obtained can be used to construct a decision support system to predict low temperatures and classify if a frost will occur. In addition, these models indicate that the DEWPT, VPD, MAXRH and MAXRAD attributes are important in the resolution of the crop frost prediction problem. As future work, the objective will be to carry out the temperature prediction with at least two hours in advance, time in which farmers can take measures to protect their crops in the case that the prediction indicates the threat of a frost. This, together with the obtaining of the local values of the attributes detected as relevant in the study, will give rise to a system that takes advantage over the systems that are currently used based on the capture of the local temperature by means of a sensor and warning farmers if a frost is occurring. Moreover in the classification task, the instances have been labeled using a simple decision rule.

As future work, these labels will be obtained through the use of clustering techniques or fuzzy/crisp discretization. In addition, for the obtaining of the models, data mining techniques that allow the treatment of imperfect data will be used. These techniques allows us to model the errors committed by the instruments for obtaining measurements

Acknowledgment

Supported by the projects TIN2017-86885-R (AEI/ FEDER, UE), TIN2016-81731-REDT, RTC-2017-6389-5 and TIN2016-78799-P (AEI/FEDER, UE), granted by the Ministry of Economy and Competitiveness of Spain (including ERDF support).

References

[1] M. C. Acock and Y. A. Pachepsky. Estimating missing weather data for agricultural simulations using group method of data handling. *Journal of Applied meteorology*, 39(7): 1176–1184, 2000.

[2] H. Al-Hiary, S. Bani-Ahmad, M. Reyalat, M. Braik and Z. ALRahamneh. Fast and accurate detection and classification of plant diseases. *International Journal of Computer Applications*, 17(1): 31–38, 2011.

[3] J. Bartok, O. Habala, P. Bednar, M. Gazak and L. Hluchỳ. Data mining and integration for predicting significant meteorological phenomena. *Procedia Computer Science*, 1(1): 37–46, 2010.

[4] C. W. Bobryk, M. A. Yost and N. R. Kitchen. Field variability and vulnerability index to identify regional precision agriculture opportunity. *Precision Agriculture*, 1–17, 2017.

[5] Fresh Plaza. http://www.freshplaza.com/article/153720/spain-frost-to-affect-80-procent-of-early-stonefruit-production, Retrieved March 10th, 2018.

[6] Fresh Plaza. http://www.freshplaza.com/article/190324/spain-about-20%2c000-tonnes-of-stone-fruit-damaged-by-frost-in-murcia, Retrieved March 10th, 2018.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1): 10–18, 2009.

[8] M. J. Ramírez, J. Hernández and C. Ferri. *Introducción a la Minería de Datos*. Pearson, 2005.

[9] IMIDA. http://www.imida.es/, Retrieved March 10th, 2018.

[10] Z. Jan, M. Abrar, S. Bashir and A. M. Mirza. Seasonal to inter-annual climate prediction using data mining knn technique. In *Wireless Networks, Information Processing and Systems: International Multi Topic Conference*, pages 40–51. Springer, 2008.

[11] C. C. Janbandhu, P. D. Meshram and M. N. Gedam. Modelling rainfall prediction using data mining method - a bayesian approach. *International Journal on Recent and Innovation Trends in Computing and Communications*, 5(3): 218–220, 2017.

[12] S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou and K. Menagias. Using data mining techniques for estimating minimum, maximum and average daily temperature values. *International Journal of Mathematical, Physical and Engineering Sciences*, 1(1): 16–20, 2008.

[13] P. Mondal, M. Basu and P. B. S. Bhadoria Critical review of precision agriculture technologies and its scope of adoption in india. *American Journal of Experimental Agriculture*, 1(3): 49–68, 2011.

[14] A. Mucherino, P. Papajorgji and P. M Pardalos. *Data mining in agriculture*, volume 34. Springer Science & Business Media, 2009.

[15] D. J. Mulla. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems engineering*, 114(4): 358–371, 2013.

[16] E. Pierpaoli, G. Carli, E. Pignatti and M. Canavari. Drivers of precision agriculture technologies adoption: A literature review. *Procedia Technology*, 8: 61–69, 2013.

[17] C. Potena, D. Nardi and A. Pretto. Fast and accurate crop and weed identification with summarized train sets for precision agriculture. In *International Conference on Intelligent Autonomous Systems*, pages 105–121. Springer, 2016.

[18] R. J. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers, 2013.

[19] C-R. Rad, O. Hancu, I-A. Takacs and G. Olteanu. Smart monitoring of potato crop: a cyber-physical system architecture model in the field of precision agriculture. *Agriculture and Agricultural Science Procedia*, 6: 73–79, 2015.

[20] Y. Radhika and M. Shashi. Atmospheric temperature prediction using support vector machines. *International journal of computer theory and engineering*, 1(1): 55–58, 2009.

[21] D. Riordan and B. K. Hansen. A fuzzy case-based system for weather prediction. *Engineering Intelligent Systems for Electrical Engineering and Communications*, 10(3): 139–146, 2002.

[22] S. K. Seelan, S. Laguette, G. M. Casady and G. A. Seielstad. Remote sensing applications for precision agriculture: A learning community approach. *Remote Sensing of Environment*, 88(1-2): 157–169, 2003.

[23] N. Sethi and K. Garg. Exploiting data mining technique for rainfall prediction. *International Journal of Computer Science and Information Technologies*, 5(3): 3982–3984, 2014.

[24] A. Stevens, Ma. Nocita, G. Tóth, L. Montanarella and B. Wesemael. Prediction of soil organic carbon at the european scale by visible and near infrared reflectance spectroscopy. *PloS one*, 8(6): 1–13, 2013.

[25] A. K. Tripathy, J. Adinarayana, D. Sudharsan, et al. Data mining and wireless sensor network for agriculture pest/disease predictions. In *2011 World Congress on Information and Communication Technologies (WICT)*, pages 1229–1234. IEEE, 2011.

[26] Y. Wang and I.H. Witten. Inducing model trees for continuous classes. In *9th European Conference on Machine Learning Poster Papers*, pages 128–137, 1997.

[27] G. Holmes, M. Hall, and E. Prank, Generating rule sets from model trees. In *In Australasian Joint Conference on Artificial Intelligence*, pages 1–12, 1999.

[28] R.J.Quinlan, Learning with continuous classes. In *5th Australian joint conference on artificial intelligence*, 92:343–348, 1992.

[29] M.H. Anisi, G. Abdul-Salaam,and A.H. Abdullah, A survey of wireless sensor network approaches and their energy consumption for monitoring farm fields in precision agriculture. *Precision Agriculture*, 16(2), 216-238, 2015.

[30] K. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, Machine learning in agriculture: A review. *Sensors*, 18(8), 2674, 2018.

[31] S. Cramer, M. Kampouridis, A.A.Freitas and A.K. Alexandridis, An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. *Expert Systems with Applications*, 85, 169–181, 2017.

[32] R.Kumar, M.P. Singh, P. Kumar and J.P. Singh, Selection Method to maximize crop yield rate using machine learning technique. *In Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*, pp. 138-145, 2015.

[33] K. Kuwata, and R. Shibasaki, Estimating corn yield in the united states with Modis Evi and Machine Learning methods. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 3(8), 2016.

[34] A.M.S.Ahamed, N.T. Mahmood, N. Hossain, M.T. Kabir,K. Das,F. Rahman, F., and R.M. Rahman, Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh. *In 2015 16th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, pp. 1-6, 2016.

Table 5. Geographical location of the weather stations used in the study. STATCODE indicates the acronym used by the station.

STATCODE	Geographical Coordinates	Altitude (m)
AL41	Lat:37 47’ 32,05” Lon:1 25’ ,39”	169
AL51	Lat:37 53’ 57,91” Lon:1 20’ 18,01”	164
AL52	Lat:37 53’ 25,64” Lon:1 18’ 34,41”	125
AL62	Lat:37 33’ 44” Lon:1 24’ 3,5”	94
AL91	Lat:37 36’ 52” Lon:1 22’ 44”	112
CA12	Lat:37 41’ 19,92” Lon:0 57’ 3,09”	30
CA21	Lat:37 49’ 53,04” Lon:1 7’ 22,49”	227
CA42	Lat:37 44’ 53,89” Lon:1 7’ 45,14”	138
CA52	Lat:37 40’ 37,25” Lon:1 4’ 14,18”	84
CA73	Lat:37 36’ 40,14” Lon:0 48’ 13,65”	92
CA91	Lat:37 41’ 56,52” Lon:1 14’ 16,96”	175
CI22	Lat:38 14’ 7,43” Lon:1 18’ 35,57”	282
CI32	Lat:38 11’ 28,96” Lon:1 15’ 28,53”	236
CI42	Lat:38 17’ 2” Lon:1 29’ 46,84”	244
CI71	Lat:38 16’ 10,2” Lon:1 35’ 6”	282
CR12	Lat:38 2’ 38,24” Lon:1 58’ 48,67”	869
CR32	Lat:38 6’ 39,35” Lon:1 40’ 59,06”	433
CR42	Lat:38 11’ 54,35” Lon:1 48’ 37,5	456
JU12	Lat:38 2’ 38,24” Lon:1 58’ 48,67”	395
JU52	Lat:38 33’ 45,57” Lon:1 6’ 44,57”	567
JU71	Lat:38 23’ 40,01” Lon:1 14’ 21,58”	401
JU81	Lat:38 19’ 11,3” Lon:1 19’ 27,58”	341
LO11	Lat:37 36’ 6,23” Lon:1 18’ 59,92”	324
LO21	Lat:37 30’ 13,86” Lon:1 41’ 38,07”	356
LO31	Lat:37 25’ 6,96” Lon:1 35’ 31,94”	31
LO51	Lat:37 29’ 16,45” Lon:1 37’ 26,29”	329
LO61	Lat:37 35’ 25,7” Lon:1 43’ 32”	319
MO12	Lat:38 0’ 25,31” Lon:1 18’ 9,23”	161
MO22	Lat:38 7’ 39,04” Lon:1 13’ 14,36”	146
MO31	Lat:38 4’ 16,1” Lon:1 14’ 1,25”	80
MO51	Lat:38 9’ 39,79” Lon:1 9’ 9,43”	197
MO62	Lat:38 6’ 47,94” Lon:1 20’ 21,94”	206
MU31	Lat:37 53’ 53,59” Lon:1 16’ 5,75”	140
MU52	Lat:37 58’ 39,15” Lon:0 59’ ,69”	125
MU62	Lat:37 56’ 24,24” Lon:1 8’ 4,99”	56
TP22	Lat:37 47’ 30,1” Lon:0 49’ 10,48”	7
TP42	Lat:37 46’ 25,89” Lon:0 53’ 54,62”	31
TP52	Lat:37 50’ 53,23” Lon:0 53’ ,75”	91
TP73	Lat:37 49’ 26,02” Lon:0 55’ 53,33”	92
TP91	Lat:37 44’ 51,81” Lon:0 59’ 12,02”	56

Agriculture and Environment Perspectives in Intelligent Systems
A. Muñoz and J. Park (Eds.)
IOS Press, 2019

Subject Index

This page intentionally left blank

A. Muñoz and J. Park (Eds.)
IOS Press, 2019

Author Index

This page intentionally left blank