A DECISION ANALYSIS MODEL FOR VEGETATION MANAGEMENT

W. Haque, J. Edwards
University of Northern British Columbia, Canada
{haque, edward2}@unbc.ca

D. Belsham
West Creek Silviculture Ltd, Canada
derek@westcreekltd.com

ABSTRACT

The application of business intelligence and analytics in non-conventional areas is an emerging research paradigm. In this paper, we present a model which uses data that has been collected over a number of years for tracking plantation and managing stand development. The data exists as legacy reports in the form of simple tables, line charts and scatter graphs which do not provide any strategic or intuitive insights. Despite the fact that this data was never collected for this purpose, we have used online analytical processing (OLAP) techniques to transform this data for querying, reporting and visual presentation on dynamic dashboards. The proposed model links regenerated stands under various operational conditions with realistic long-term growth and yield predictions.

Keywords: Business Intelligence, Silviculture, Vegetation Management

1 INTRODUCTION

Business intelligence (BI) provides a comprehensive high-level analysis which can empower users to quickly assess trends in a vast quantity of data, while still being able to provide views of low level data through drill-downs. BI tool stacks are becoming more and more common in commercial uses for analyzing traditional business data; some suites provide templates for basic reporting and analysis, further simplifying the overall process. In this paper, we extend the concept to a somewhat non-traditional application of BI, that is, silviculture. Generally, silviculture is the science of growing trees, but in a number of contexts this term is also associated with the processes following a logging operation to restore the site back to its original state. The silviculture forester is mandated by law to reforest a logged area within specified time limits. This requires consideration of several factors when planning a reforestation project including identification of the tree species to plant, the plantation density, the likelihood of brush competition, and the required pre and post planting treatments. These factors are also related to environmental factors such as soil type, predominant sunlight, altitude above sea level, and moisture content. In sum, the selected problem definitely suggests a multi-dimensional approach for data analysis.

Typically, when data is collected with well-defined analytical goals, the data aggregation across various dimensions is relatively straightforward. On the other hand, when the data is collected for legacy reporting, it poses serious challenges for analytical processing. In this paper, we use a standard set of tools with a non-standard set of data to determine the benefits of BI in a scenario where it would typically not be applied. The data source used for this study is the NIVMA (Northern Interior Vegetation Management Association) database, which relates to silvicultural operations in the western provinces of Canada [1]. The NIVMA data was collected over several years by creating permanent plots in representative ecosystem sites throughout the provinces. The various sites include samples of different tree species, site conditions, brush competition, and site treatments, if any. Among other things, the individual tree data includes periodic measurement of the height, health, and root collar diameters.

From data warehousing perspective, the NIVMA database represents low quality and noisy data containing incomplete, incorrect, inconsistent and duplicate values. Thus, before performing any analytics, it was necessary to transform this data into high quality data and then load it into a data warehouse (DW). The analytical and reporting layers of BI software are then used to provide the analysis that the NIVMA database was originally intended for. Besides displaying the desired features within specific “ecozones” as trends through charts and other data visualization techniques, the real benefit of using BI is in comparative and predictive analysis. For example, a user can use the generated reports to compare the yield of a planted block with results from blocks which share features such as brush density, climate, and tree species. This allows the user to identify potential issues and underlying causes early in the process. Further, a user could view all blocks by average height in order to examine overall trends, or just the blocks which share the same ecozone as the block with the lowest average height. This information can then be used to determine any noteworthy differences in height or other metrics which could imply a correlation to ecozone. Such comparisons are invaluable for future block planting.

2 RELATED WORK

Currently, there is little if any work being done to apply business intelligence solutions to natural resource based industries. In particular, we are not aware of any models that are based on OLAP cubes. There are, however, several systems available to analyze forestry resource information providing visualization and analysis in a different manner [2]. One such system is the “Table Interpolation Program for Stand Yields” or TIPSY [3]. TIPSY originated several decades ago and has been under continuing development since then. The purpose of TIPSY is to interpolate yield tables and customize the information therein so that a visualization and summary of the data can be displayed. Generally, this is used to simulate a specific set of operational conditions in order to project growth within a single stand. Another system which is in active use is the “Tree and Stand Simulator”, or TASS [4]. TASS has a purpose similar to that of TIPSY, which is to generate growth and yield information [5]. However, the information generated is limited to modeling of purely coniferous stands of even-age. TASS is, in fact, used to generate the yield tables which are used in TIPSY, and is reliant on several sub-models itself.

A system closely related to TIPSY and TASS is SYLVER (Silvicultural Treatments on Yield, Lumber Value, and Economic Return) [6]. SYLVER predicts wood quality, yield, and general financial return for stands using TASS and other subsystems, but is limited to several specific species, having full functionality available only for Douglas fir.

3 TEST ENVIRONMENT

We have chosen to use Microsoft’s BI tool stack for this study due to academic licensing and ease of integration with other commonly used applications. Microsoft’s Business Intelligence Development Studio [7] includes SQL Server Management Studio (SSMS) for storing databases, SQL Server Integration Services (SSIS) for developing the data warehouse, SQL Server Analysis Services (SSAS) for building a cube structure on top of data warehouse, and SQL Server Reporting Services (SSRS) for reporting and visualization. The original data was stored in a Microsoft Access database. We used Microsoft's migration assistant to convert it into an equivalent SQL Server database for compatibility with the chosen tool stack.

4 NIVMA AND THE PROPOSED BI MODEL

The purpose of the NIVMA database was to develop growth models for trees belonging to specific ecosystems. This allows for better, educated decisions regarding forest restocking that ultimately yield a greater return than previously possible. For instance, with proper analysis, it is possible to relate annual growth by species to various site conditions. The analytics can be used to identify species that perform best in a particular environment, or to suggest desirable post planting activity. The cube can also be used for scenario testing, that is, to determine what would happen if we applied regime A as opposed to regime B. Similarly, it could be used to track the effectiveness of a suggested treatment that may have been applied to a poorly performing site.

A data cube is constructed in a way which is conceptually similar to a physical cube. The dimensions of the cube represent metrics along which certain performance indicators are to be measured; the data related to the indicators itself fills the volume of the cube. The online analytical processing (OLAP) capability of such cube structure allows us to “slice and dice” the data along the dimensions to view specific segments of the metric data. For example, if we take *time* to be a dimension and *sales* to be a performance metric, we could slice the cube to view sales numbers for only the year 2010. Unlike a physical cube, the OLAP cube data structure can have many more than three dimensions and a varying number of metrics. As such the data warehouse must be built with this in mind so that work in SSAS focuses on items such as establishing foreign key relationships, obtaining calculated measures, and creating attribute hierarchies.

The first step in creating the data warehouse is to identify dimensions of the cube, and specify the metrics or facts. The dimensions chosen for this data warehouse are: 1) Ecozone, 2) Time, 3) Soil, 4) Species, 5) Location, and 6) Damage. The warehouse is built using the star schema [8], which means that each of these dimensions occupy a table which is linked to the two fact tables in the data warehouse through foreign key relationships. The fact tables contain important facts such as height and crown radii, and foreign keys such as measurement dates and species which can be used to establish a relationship with the dimension tables for each of the individual trees and stands, respectively, for which we have repeated measurements.

Every Block in the NIVMA database is described as a set of Quads which can then be broken down to a specific Grid. The Quads and Grids are used to describe the location of each measurement within a Block and is one of the key ways to tie the different measurements together. There are two key issues that occur with the original database when we try to associate repeatedly measured trees with density data. The first issue that occurs is that, in a few cases, the Grids and Quads are not associated properly for either the density or tree data. The second, more serious, problem is that not every density data entry has associated repeatedly measured tree values and the same is true in the other direction. The consequence of this is that when we try to create a regular relationship between these two tables based on a Grid/Quad key there are instances where no foreign key exists. We solved this problem by inserting null records into Grids and Quads tables where necessary in order to ensure that the relationship exists. This has resulted in adding 2150 null records to a table that consists of 36,252 records. The other challenges posed by the original data resulted from inconsistent procedures and specifications for data collection. For instance, there is no standard for the starting age of a tree, number of measurements for a tree, lack of standard units of measurement, and ways to specify, for instance, a dead tree. Each of the problems encountered was dealt with in a way that ensured data integrity at all times. No changes to the underlying data were made which could cause significant alterations to the final outcome. The final reports, therefore, reflect results based purely on the original un-altered data.

4.1 Defining Key Performance Indicators

We have identified two primary indicators to provide the high level performance analysis that identifies the best or worst performing groups of trees quickly. First, we want to be able to identify the rate at which a tree is growing, aggregated to the level of a stand or species. The second indicator is the number of trees which are likely to survive based on the final measurement for each tree. One of the key factors by which we want to be able to slice both of these values is the *ecozone*, which is a combination of zone, subzone and variant. We also want to slice the measurements by the species of the existing trees. A single stand is capable of having multiple ecozones analysis of the growth trend, we aggregate the same KPIs in a slightly different way depending on whether the trend for individual trees or at the stand level is needed. In addition, the user should be able to view that data sliced across more dimensions at the top level that could potentially reveal interesting trends and provide drill down capabilities.

The growth rate of an individual tree can be determined using the current and previous year's data. However, we want a single value to represent a specific stand. The nature of the data presents several problems in building this value by simply aggregating the sum of heights from the tree measurements. The first problem is the sporadic nature of the individual measurements. The stands do not contain an even distribution of measurements and the number of measurements per tree varies even within the same stand. A stand has the potential to contain sets of trees that are measured over different periods of time; for instance, the first set of measurements could begin in 2000 and end in 2005, while the second set may begin in 2005. This results in having both very large and very small trees in the same year. Yet another issue with the height data is that some trees are measured in centimeters while others are measured in meters. To solve these problems we create a calculated measure in the cube which computes the growth rate as a percentage at the bottom level. At the tree level, the growth rate is specified as a percentage by examining the current height of a tree versus the height for the previous year. At the stand level, an average of those percentages is taken thus providing the overall growth rate for that stand. This calculated measure thus returns similar values regardless of the base measurement unit.

In order to calculate the Survival Rate of a Stand, we have two possible data fields to indicate that a tree has died. The first possibility is that a tree has a final height of zero. This, however, does not give us the whole picture, as a tree can still have a valid height even though it is not likely to survive. A second statistic, *vigour*, is provided with each measurement. Vigour has a range of one to three, where three indicates that the tree is strong and healthy and one indicates that the tree is dead or will not survive. In order to calculate the survival rate, we have used the following formula:

$$\text{Survival Rate} = \frac{\text{Final Vigour}}{3 * \#\text{ of unique Trees}} \qquad (1)$$

This gives us a value that indicates a 100% survival rate if each tree has a final vigour of 3 and a lower % as each tree does worse. This measurement works nicely for both tree and stand level analyses.

4.2 Defining the Low Level Analysis for Stands

Once a key Stand has been identified, we provide low level analysis to analyze the performance of a specific stand. Some of the key factors involved in a tree's growth rate include damage that has occurred, the density of the stand, the average height of the trees by year, the stand's treatment history and the Comeau competition index [9]. Each of these factors resides in a different table within the NIVMA database and must be linked to the stand data. The Comeau competition index gives us a measure of how the tree is doing compared to the surrounding brush (*brush competition).* It is generated by the following formula:

$$\text{Comeau index} = \frac{\%\text{ cover of brush} * \text{height of brush}}{\text{average height of trees}} \qquad (2)$$

This produces a single value that indicates the overall health of the stand and can be shown on a yearly basis. The density of a stand is measured as the average number of stems planted multiplied by the size of the plot in square meters. In NIVMA, density data is collected on a semi-yearly basis at the quad level. There are multiple measurements per quad each year for the same species.

Calculating the average tree height by year becomes much easier when we drill down to a specific stand. When the block is considered as a whole, there are still issues that arise as trees in different quads usually have different measurement schedules. However, trees measured within the same quad level generally seem to be measured at the same time, which makes this the desirable level at which the analysis should be provided. Aggregating at this level also works nicely to reduce the amount of data provided to the user as there are still about 36 trees measured per block which quickly clutters any kind of graph or chart.

Each stand in NIVMA has some historical data associated with it which indicates the type and date of the treatments (for instance, spraying herbicides) applied to the stand. This dimension is particularly easy to deal with because the goal is to provide the user with a history of what has happened in that specific stand; aggregations are not relevant here. The end user is able to use this history to understand the rest of the report associated with that block.

5 CUBE DESIGN & DATA INTEGRATION

The design of the OLAP cube determines the data integration process as each of the data warehouse tables must correspond to one of either the dimension or fact tables identified in the design phase. Consequently, our integration process consisted of two major steps: the population of the six dimensions, and the two fact tables.

Before these two steps are undertaken, there is a small preprocessing step to ensure that all tables are cleared of data before the process is run. This is necessary because the underlying data is not regularly updated, but only replaced as a batch. Thus it would not be desirable to repeat any records by mistakenly adding rows of data which overlap with a previous batch.

After the data warehouse has been emptied, we begin to aggregate data from approximately 270 tables in the NIVMA database and feed it into each of the relevant dimension tables in parallel. This is largely uncomplicated, although in some cases in order to maintain the integrity of the schema some changes must be made to the data after it is loaded. For example, there are several records for which no ecozone data is available; in these cases a record is inserted into the ecozone dimension with null attributes so that the foreign key relationships with the fact tables can still exist.

The final step is to populate the fact tables. The stand level and tree level data is populated in sequence, beginning with the more aggregate data in the stand level fact table. The process for each of these fact tables consists largely of using merge-join transformations to combine the relevant set of incremental measurement tables from the source data. In order to have a common key to use for these joins, a surrogate key is created using a concatenation of attributes which forms a candidate key for the results of the join. After this series of joins is completed to the point that all requisite data has been included, foreign keys to each of the dimensions are derived where required. Once keys are in place for the relationships in the schema, the cube can be processed for analysis. The cube design and the complete Extract-Transform-Load (ETL) process is described in [10].

6 DATA CHALLENGES

Whenever work is done with non-business data, unique challenges present themselves during the process of converting this data to a state suitable for analysis. Specifically with NIVMA, the largest issue lies within the consistency of the data. As the data was collected by several different agencies, there are somewhat differing standards for data collection and quality. An issue stemming from this is the inconsistency with which measurements are taken in terms of regularly spaced periods of time. Even more issues arise in identifying the age of individual trees as the age of a tree at the time of initial measurement is not recorded in a standardized way. The age identification of trees is especially important for this analysis as it will correlate directly to many of our important metrics such as growth rate and damage. Largely, the problems associated with data being recorded heterogeneously by these multiple agencies are analogous to those which would be experienced by an organization integrating more standard data from multiple business sources such as data marts for differing departments.

One important example of the lack of standardization in measurement can be seen in the initial measurements of individual trees. The majority of repeatedly measured trees have an initial height measurement which is less than 100 while some trees have an initial height which is greater than 1000, each in an unspecified unit of measurement. Generally, the larger trees exhibit variance by a factor of roughly ten, which has led us to the assumption that the heights of larger trees were recorded using a different unit of measurement. Such a significant variance creates obvious problems for analysis, for example if examining the average height of all trees within a certain species is valuable, this skewed average could lead to misinformed decision making at the end user level. A second issue with the height measurements is the way in which a dead tree is represented. In most cases a value of zero indicates a dead tree, but in a few cases a dead tree has been indicated with a value of 9999; this once again can potentially skew any aggregation. We solve the first issue by making the assumption that the different heights occur in either cm or mm, and we scale the size of a tree accordingly. Approximately 0.5% of trees in the database appear disproportionately large, and thus the overall effect of this change should only be strongly apparent in specific tree groupings (such as species) which are dominated by these very large trees. The dead tree issue is resolved in a similar manner, by simply setting the trees with a 9999 height to a value of zero.

Other issues stem not from the way in which the data was collected but in the way the NIVMA database was structured. Because the database was intended for legacy reporting and is built using a relational schema, the table structures therein do not lend themselves to dimensional analysis. In fact, the roughly 200 tables in the original NIVMA database are transformed into only 13 tables for the data warehouse which provides the framework for the

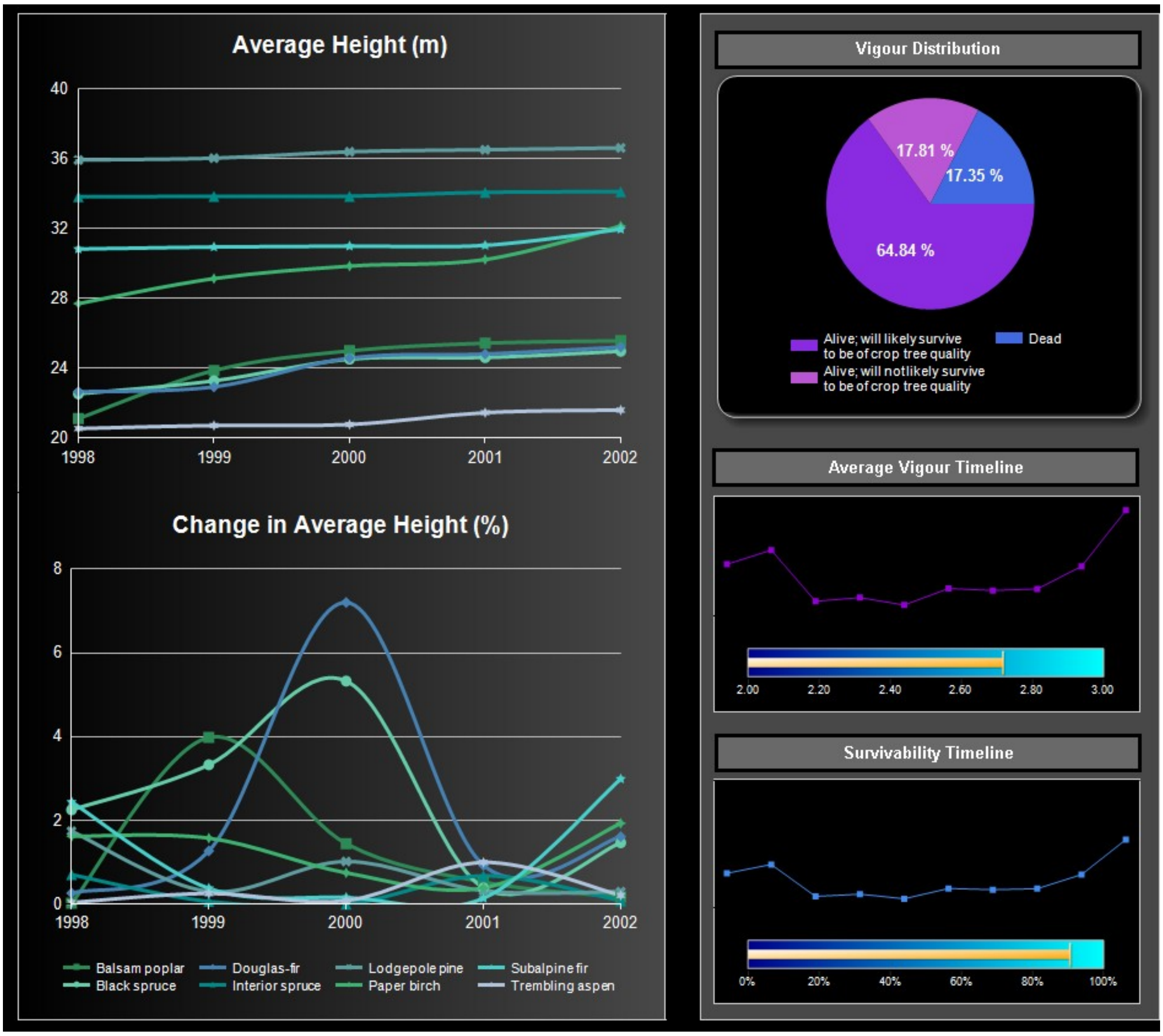
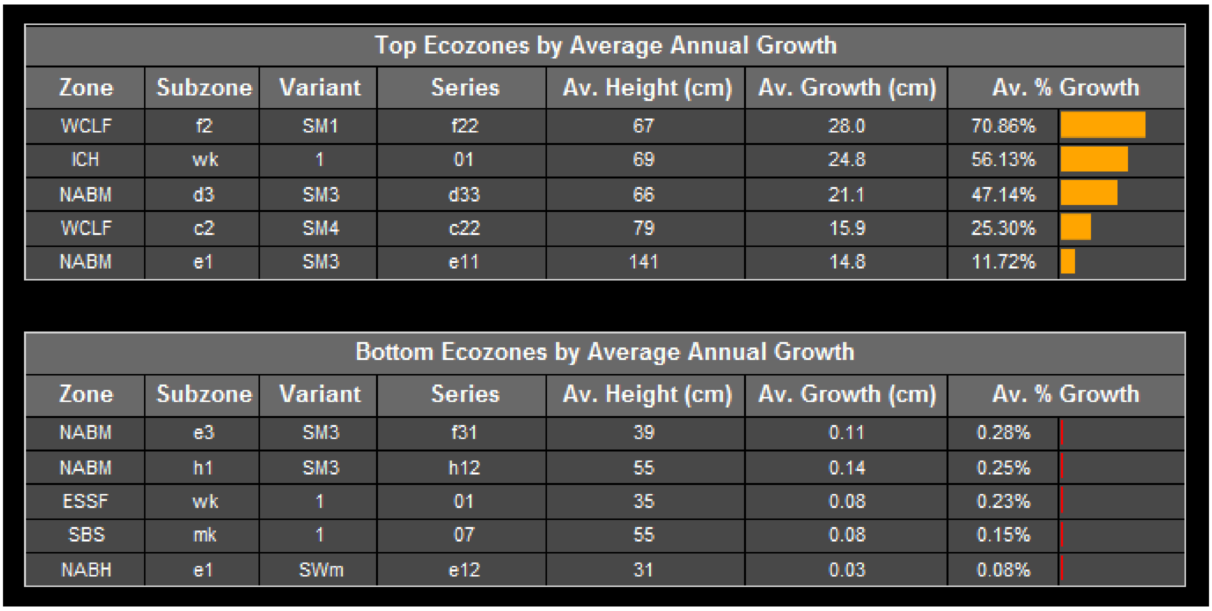


Figure 1: Sample dashboard for vegetation management



Top Ecozones by Average Annual Growth

Zone	Subzone	Variant	Series	Av. Height (cm)	Av. Growth (cm)	Av. % Growth
WCLF	f2	SM1	f22	67	28.0	70.86%
ICH	wk	1	01	69	24.8	56.13%
NABM	d3	SM3	d33	66	21.1	47.14%
WCLF	c2	SM4	c22	79	15.9	25.30%
NABM	e1	SM3	e11	141	14.8	11.72%

Bottom Ecozones by Average Annual Growth

Zone	Subzone	Variant	Series	Av. Height (cm)	Av. Growth (cm)	Av. % Growth
NABM	e3	SM3	f31	39	0.11	0.28%
NABM	h1	SM3	h12	55	0.14	0.25%
ESSF	wk	1	01	35	0.08	0.23%
SBS	mk	1	07	55	0.08	0.15%
NABH	e1	SWm	e12	31	0.03	0.08%

Figure 2: A drill-down report for ecozone analysis

OLAP cube. Each of the dimension tables which are used for analysis is constructed through a sequence of merge and lookup operations which can involve up to ten smaller tables from the source. The actual process of joining these tables poses some interesting problems in terms of forming primary keys and avoiding lossy joins while using the pre-existing keys. Ultimately, the keys which were derived in most of the dimension tables are concatenations of metrics which were each represented by their own tables previously; for example zone, subzone, and variant are combined to form a key for the ecozone dimension. This derived ecozone key is a string which is used and well understood by silviculturalists and foresters in real world situations.

In the NIVMA database, individual trees are measured on a semi-yearly basis; however, the year of a first measurement in each block is not consistent amongst all blocks. A further complication occurs in that there is a large variance between the sizes of first height measurements for trees which seems to indicate that they may not all be measured after the same period of growth. At a specific block level this does not create problems as the vast majority of blocks have a consistent measurement pattern. At a high level however, we want to be able to aggregate all of the growth data for a species to determine the typical growth patterns. These high level growth trends can then be used to identify specific blocks that are performing poorly or exceptionally, and identify the individual factors that may have led to these growth patterns.

7 QUERYING AND REPORTING

The reporting and data visualization aspect of this study is where the power of a business intelligence solution becomes apparent. A dashboard has been created to show the features offered by high level analysis when sliced along different dimensions.

Firstly, at the top left of Fig. 1, a line graph displays the average height of trees in a stand. The parameters of this chart have been set to display data for the range from 1998 to 2002 utilizing the time dimension, while displaying data for eight selected species utilizing the species dimension. These parameters are selectable from drop-down menus which are not shown here due to space considerations. This chart enables us to view the general trends for height and thereby compare the average lumber volume offered by each of the species in comparison to one another. The chart also allows us to identify any patterns of change in average height which may be indicative of problems, or spurts of growth which may have been caused by an identifiable factor. Likewise, a chart for average change in height (growth value) is displayed. This chart is also parameterized, but the trends shown provide a different perspective of changes in the average height.

Along the right side of Fig. 1, there are several charts displayed which relate to the vigour value of stands. The pie chart, representing the vigour distribution of a stand, provides a breakdown of percentage of trees in a stand for each of the three distinct vigour categories. This gives a somewhat different picture than simply taking an average of the vigour seen in a group of trees. Following this, two sparklines are shown, each accompanied by a gauge. The first set displays the average vigour of a set of trees. This analysis allows the user to see which vigour rating a group of trees is nearer to on average, whether healthy or unhealthy. As opposed to the distribution shown in the pie chart, this gives a less detailed view but is easier to make inferences from. The sparkline shows the general trend in vigour over the past ten years without specifying values explicitly, while the gauge below allows the user to view the exact value for average vigour in the current year in order to gain a sense of context for the trend. In the actual application, the tooltips allow for popping up actual values on mouse-over hover. The second set, containing the sparkline and gauge for survivability is structurally very similar. This visualization allows the user to view the overall likelihood of a tree dying, as opposed to simply the likelihood that a tree will achieve crop quality during its growth.

Fig. 2 shows a more specific analysis of the data, giving a list of the top and bottom five ecozones compared to previous year. This chart is also parameterized to allow for specification of the species of interest. The information contained in the table is very useful for considerations in future planting. The overall height of the trees listed is shown, as well as the growth over the previous year as a measured quantity and then as a percentage which is accompanied by a bar chart which shows the trend in the five ecological zones. The ecozones are displayed as columns broken into the zone, subzone, and variant which describe the plot in which a tree was grown as well as the site series descriptor for the location. This allows for identification of environmental factors which may have contributed to the success or general lack of growth seen in the trees. For example, we observe that the ‘NABM’ zone is common amongst the ecozones with poorest growth, though it also appears in some of the ecozones with the greatest growth rate; this could mean that the zone was not the only factor relevant to the growth. In each of the NABM zones seen in the poorly performing ecozones the variants are all ‘SM3’, whereas the site series and subzones are dissimilar. This would indicate that perhaps the variants and zones are not the factors which caused a gap in the growth rates of the trees shown, but that the site series’ and subzones warrant further investigation, as provided in the drill-down reports.

The scope of this paper is to demonstrate the use of business intelligence analytics to non-conventional areas; thus, reports which are too specific to silviculture have not been included here.

8 CONCLUSIONS

In this paper, we have demonstrated how OLAP techniques can be applied to non-conventional applications. In particular, we used a database which contained useful, but incoherent data that was collected without BI analytics in mind. The data was extracted, transformed and loaded into a data warehouse. Specifically, the data was examined using OLAP techniques resulting in very intuitive, useful analytics and reports. We have provided examples of some of the possible reports and demonstrated how they could be effectively used in the decision making process of silviculture in order to achieve maximum output from stands and individual trees in terms of lumber volume. Currently, there is little work being done to apply business intelligence to fields such as vegetation management, but given the evidence of its benefits as shown in this paper, there will be more development in this area in the future.

9 REFERENCES

[1] Stephen M Smith: "NIVMA's New Models of Regenerated Forest Stands", Victoria, Forest Science Program Project No. Y051235 2005.

[2] C. Mario Di Lucca: "TASS/SYLVER/TIPSY: Systems for predicting the impact of silvicultural practices on yield, lumber value, economic return and other benefits," in *Stand Density Management Conference: Using the Planning Tools*, Victoria, 1999, pp. 7-16.

[3] Ministry of Forests and Range Research Branch: (2007, November) Table Interpolation Program for Stand Yields. [Online]. http://www.for.gov.bc.ca/hre/gymodels/tipsy/index.htm

[4] Ministry of Forests and Range Research Branch: (2002, October) Tree and Stand Simulator. [Online]. http://www.for.gov.bc.ca/hre/gymodels/TASS/index.htm

[5] C. Mario Di Lucca, Jim Goudie, and Steve Stearns-Smith: "TASS III: A new generation growth and yield prediction model for complex stands in British Columbia," in *World Forestry Congress*, Buenos Aires, 2009.

[6] Ministry of Forests and Range Research Branch: (2002, October) Silvicultural Treatments on Yield, Lumber Value, and Economic Return. [Online]. http://www.for.gov.bc.ca/hre/gymodels/SYLVER/index.htm

[7] Microsoft: (SQL Server 2008 R2) Introducing Business Intelligence Development Studio. [Online]. http://msdn.microsoft.com/en-

us/library/ms173767.aspx

[8] Ralph Kimball and Margy Ross: *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*, 2nd ed. Canada: John Wiley & Sons, 2002.

[9] P G Comeau, T F Braumandi, and C Y Xie: "Effects of overtopping vegetation on light availability and growth of Engelmann spruce (Picea engelmannii) seedlings," *Canadian Journal of Forest Research*, vol. 23, no. 10, pp. 2044-2048, October 1993.

[10] Waqar Haque and Jacob Edwards: "High Quality Analytics with Poor Quality Data - A Case Study in Silviculture," in *Proc of Int'l Conference on Business Intelligence and Data Warehousing*, Singapore, 2011, pp. B1-B7.