# NGHIÊN CỨU ỨNG DỤNG CÔNG NGHỆ THÔNG TIN ĐỂ PHÂN TÍCH, THỐNG KÊ CƠ SỞ DỮ LIỆU NGUỒN GEN LÚA THUỘC DỰ ÁN PHÁT TRIỂN NGÂN HÀNG GEN CÂY TRỒNG QUỐC GIA, 2011-2015

Vũ Đình Tú[1], Nguyễn Thị Hiền[1], Nguyễn Chí Tín[1], Nguyễn Tiến Hưng[1]

## TÓM TẮT

Tin học hóa nền nông nghiệp được coi là cuộc cách mạng xanh ở thế kỷ 21. Trong nông nghiệp nói chung và công tác bảo tồn tài nguyên thực vật phục vụ nông nghiệp nói riêng thì Công nghệ thông tin (CNTT) không chỉ là phương tiện hỗ trợ mà có thể trở thành lực lượng lao động quan trọng. Vì vậy, việc ứng dụng công nghệ thông tin để phân tích, thống kê cơ sở dữ liệu là một trong những công việc rất quan trọng của công tác bảo tồn TNTV của Trung tâm Tài nguyên thực vật. Để khai thác cơ sở dữ liệu Dự án phát triển Ngân hàng gen cây trồng quốc giai giai đoạn 2011-20150, Bộ môn Dữ liệu và Thông tin TNTV đã ứng dụng hiệu quả một số phần mềm như Excel, SPSS, QGIS, Infographic… để làm sạch dữ liệu, phân tích, thống kê, trình bày sơ sở dữ liệu nguồn gen thu thập được (Bao gồm cơ sở dữ liệu về thông tin lai lịch và mô tả đánh giá nguồn gen). Báo cáo này chủ yếu giới thiệu một số khái niệm và kết quả ứng dụng CNTT vào bảo tồn tài nguyên thực vật phục vụ nông nghiệp của Bộ môn trong thời gian qua nhằm giúp cho các nhà quản lý, các cán bộ nghiên cứu trong Trung tâm hiểu rõ hơn các hoạt động cứu và phục vụ nghiên cứu của Bộ môn hiện nay.

*Từ khóa:* Công nghệ thông tin; Phân tích, thống kê dữ liệu; Lúa (*Oryza sativa* L.)

## I. ĐẶT VẤN ĐỀ

Dự án phát triển Ngân hàng gen cây trồng Quốc gia giai đoạn 2011-2015 do Trung tâm Tài nguyên thực vật thực hiện đã thu thập được 12.758 mẫu giống của 119 loại cây trồng trên toàn quốc. Trong đó, Lúa (*Oryza sativa* L.) là loại cây trồng thu thập được nhiều và đa dạng với số lượng 1.704 mẫu nguồn gen.

Hoạt động bảo tồn và sử dụng bền vững quỹ gen cây trồng đòi hỏi quá trình thu thập, lưu trữ thông tin và sinh ra một lượng dữ liệu khổng lồ. Chính vì vậy, việc xây dựng hệ thống cơ sở dữ liệu có khả năng cung cấp dữ liệu có độ tin cây cao cho nhiều đối tượng sử dụng là một công việc không thể thiếu của hoạt động bảo tồn. Hiện tại, cơ sở dữ liệu của Trung tâm Tài nguyên thực vật bao gồm dữ liệu Lai lịch, Mô tả đánh giá ban đầu, Mô tả đánh giá chi tiết, Hình ảnh…được cung cấp từ các hoạt động bảo tồn. Đến nay đã có hàng triệu trường dữ liệu cho các Loại cây trồng khác nhau. Khối lượng dữ liệu ngày càng nhiều dẫn đến việc lưu trữ và phân tích, thống kê dữ liệu sẽ gặp phải những khiếm khuyết

[1] Bộ Môn Dữ liệu và Thông tin TNTV

nhất định. Trước kia, việc nhập dữ liệu, thống kê dữ liệu thường được tiến hành thủ công và được đối soát theo bản mẫu gây mất rất nhiều thời gian, tiền của và công sức. Từ khi áp dụng công nghệ thông tin trong khâu xử lý và tổng hợp số liệu thống kê, thời gian xử lý và tổng hợp cho một cuộc điều tra được rút ngắn đáng kể. Hơn thế nữa, sử dụng các chương trình máy tính trong khâu xử lý và tổng hợp số liệu còn cho phép nâng cao được chất lượng số liệu thống kê thống qua các chương trình kiểm tra logic và sửa lỗi. Bài báo cáo đưa ra các khái niệm, công cụ hỗ trợ, phần mềm chuyên ngành để có thể giúp ích trong công tác tiền xử lý dữ liệu, phân tích, thống kê cơ sở dữ liệu nguồn gen Lúa thu thập và mô tả, đánh giá trong dự án phát triển ngân hàng gen cây trồng quốc gia (2011-2015)

## II. VẬT LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

- Dựa vào cơ sở dữ liệu thông tin nguồn gen lúa đang được quản lý tại Bộ môn Dữ liệu và Thông tin tài nguyên thực vật bao gồm dữ liệu thu thập nguồn gen (nhóm dữ liệu Đăng kí, Lai lịch), dữ liệu mô tả đánh giá ban đầu nguồn gen (nhóm dữ liệu Mô tả, đánh giá nguồn gen), chúng tôi chọn ra bộ cơ sở dữ liệu của 1.704 nguồn gen lúa được thu thập bởi Dự án Phát triển ngân hàng gen cây trồng quốc gia và bộ cơ sở dữ liệu của 940/1.704 mẫu giống đã được tiến hành mô tả, đánh giá đặc điểm nông sinh học ban đầu.

- Từ bộ cơ sở dữ liệu nguồn gen lúa, chúng tôi tiến hành tiền xử lý dữ liệu bằng phương pháp làm sạch dữ liệu (data cleaning).Từ nguồn dữ liệu đã được xử lý đó chúng tôi tiến hành phân tích, thống kê nguồn gen lúa theo vùng sinh thái, theo nguồn gốc dân tộc sở hữu, theo dữ liệu mô tả đánh giá một số các chỉ tiêu cơ bản để tổng hợp các bảng số liệu, thông tin.

### 2.1 Ứng dụng Làm sạch dữ liệu (Data Cleaning) để rà soát lại dữ liệu, nhằm đảm bảo rằng các dữ liệu đều đồng nhất và chính xác ở mức độ cao nhất.

#### 2.1.1. Kiểm tra, chuẩn hóa giá trị dữ liệu:

- Quy trình kiểm tra, chuẩn hóa giá trị dữ liệu được tiến hành trên Nhóm dữ liệu Đăng ký, dữ liệu Lai lịch , dữ liệu Mô tả, đánh giá. Quy trình này thực hiện trên các dữ liệu Dữ liệu chính tả (Số đăng kí, Tên mẫu nguồn gen); Dữ liệu địa lý (Tỉnh, huyện, xã); Dữ liệu tọa độ (Kinh độ, Vĩ độ); Dữ liệu dân tộc; Dữ liệu thời gian (Ngày/tháng/năm)

- Tất cả các công đoạn được tiến hành trên tệp (file) Excel. Để đảm bảo an toàn dữ liệu trong khi thao tác chúng tôi tạo bản sao lưu dữ liệu ban đầu trong một file làm việc khác.

- Các bước chung cho thao tác một trường dữ liệu là:

- Chèn một cột mới (B) bên cạnh cột gốc (A) cần làm sạch.

- Thêm công thức sẽ biến đổi dữ liệu ở trên cùng của cột mới (B).
- Điền công thức trong cột mới (B). Trong bảng Excel, một cột được tính toán tự động được tạo bằng giá trị điền xuống dưới.
- Chọn cột mới (B), sao chép nó, sau đó dán dưới dạng giá trị vào cột mới (B).
- Loại bỏ cột gốc (A), chuyển đổi cột mới từ B đến A.

- Loại bỏ khoảng trắng và các ký tự thay thế, chỉnh sửa chính tả: sử dụng một số hàm trong tệp Excel như Find & Replace, TRIM. VLookup…

- Chuẩn hóa dữ liệu về địa giới hành chính (tỉnh/huyện/xã) của các nguồn gen bằng cách đối chiếu với cơ sở dữ liệu chuẩn về địa giới hành chính

- Chuyển đổi dữ liệu tọa độ (Kinh độ, Vĩ đô) đồng nhất về hệ tọa độ Decartes (hệ tọa độ không gian 2 chiều bằng cặp số tọa độ x, y). VD: Chiềng Sại, Bắc Yên, Sơn La có tọa độ (Kinh độ, Vĩ độ) Decartes là: (104.506667, 21.069722)

- Chuẩn hóa dữ liệu về dân tộc của các nguồn gen bằng cách đối chiếu với cơ sở dữ liệu "54 dân tộc Việt Nam" của Ủy ban dân tộc Việt Nam.

- Chuẩn hóa dữ liệu thời gian về định dạng ngày tháng năm (dd/mm/yyyy) (VD: 11/09/2014)

**2.1.2. Nhận diện, xử lý phần tử ngoại lai (outliers) và giảm thiểu nhiễu (noise data)**

- Xác định phần tử ngoại lai bằng một số phương pháp: phân bố thống kê (statistical distributionbased), khoảng cách (distance-based), phương pháp giảm thiểu nhiễu phân cụm (clustering) ...để hiệu chỉnh dữ liệu

**2.1.3. Nhận diện, xử lý dữ liệu bị thiếu (missing data)**

- Sử dụng phần mềm thống kê số liệu SPSS Statistics để xác định được các giá trị bị thiếu (missing values) và qui đổi giá trị thiếu về hằng số chung.

**2.2. Ứng dụng Hệ thống thông tin địa lý (Geographic information system- GIS) để phân tích, thống kê dữ liệu không gian (dữ liệu bản đồ) của các mẫu nguồn gen**.

- Dựa vào dữ liệu về tọa độ (kinh độ, vĩ độ) được chuẩn hóa theo Hệ tọa độ Decartes và sử dụng hệ tọa độ quốc tế WGS 84 trên GIS chúng tôi bước đầu ứng dụng QGIS (Window, Mac OS X Linux) trên lớp bản đồ nền 63 tỉnh thành Việt Nam để thống kê phân bố nguồn gen Lúa được thu thập trên toàn quốc và 8 vùng sinh thái nông nghiệp

**2.3. Ứng dụng phần mềm xử lý số liệu Excel, IBM SPSS Statistic để phân tích, thống kê dữ liệu lai lịch, dữ liệu mô tả đánh giá nguồn gen.**

### 2.4. Ứng dụng Infographic (Information graphic) (Adobe Photoshop, Adobe Illustrator) để đồ họa trực quan thông tin, dữ liệu nguồn gen Lúa

## III. KẾT QUẢ VÀ THẢO LUẬN

### 3.1 Ứng dụng Làm sạch dữ liệu (Data Cleaning) để rà soát lại dữ liệu

Làm sạch dữ liệu (Data cleaning) là công việc hết sức quan trọng trong quá trình tiền xử lý dữ liệu để đảm bảo tính chính xác (accuracy), tính hiện hành (currency), tính toàn vẹn (completeness), tính nhất quán (consistency). Một thuật ngữ về chuyên ngành dữ liệu được đưa ra đó là "garbage in, garbage out" (dữ liệu đầu vào là rác thì dữ liệu đầu ra sẽ là rác). Nếu chúng ta cung cấp một tập dữ liệu chứa thông tin rác, thì kết quả cuối cùng chúng ta nhận được cũng sẽ là rác. Do đó, khi nhận được một lượng lớn dữ liệu, việc đầu tiên mà chúng ta cần nghĩ đến là tiền xử lý tập dữ liệu đó, để có thể hạn chế rác (garbage) và sử dụng chúng để khai phá sau này. Kết quả làm sạch bằng một số phương pháp:

#### 3.1.1 Kiểm tra, chuẩn hóa giá trị dữ liệu

Kết quả kiểm tra, chuẩn hóa dữ liệu được trình bày tại Bảng 1:

**Bảng 1**: Thống kế số lượng dữ liệu được kiểm tra, chuẩn hóa giá trị

| Loại dữ liệu | Dữ liệu chính tả | Dữ liệu địa giới hành chính (Tỉnh/Huyện/Xã) | Dữ liệu tọa độ | Dữ liệu dân tộc | Dữ liệu thời gian |
|---|---|---|---|---|---|
| **Số dữ liệu cần kiểm tra** | 1704 | 39/136/407 | 1704 | 48 | 1704 |
| **Số dữ liệu được hiệu chỉnh** | 22 | 39/135/407 | 168 | 32 | 0 |

- Trong quá trình kiểm tra, chuẩn hóa chúng tôi nhận thấy Dữ liệu địa giới hành chính hầu như chính xác tuyệt đối (chỉ có 1 huyện Gia Nghĩa bị nhập liệu sai nên thành 2 huyện); Dữ liệu chính tả bị sai chủ yếu bởi chỉ tiêu Tên nguồn gen bị lỗi kí tự dấu cách; Dữ liệu tọa độ thì có 168 tọa độ được chuyển đổi từ hệ tọa độ GPS về hệ tọa độ Decartes, dữ liệu dân tộc chủ yếu bị lỗi khi cán bộ thu thập viết tên sai; Dữ liệu thời gian có tỷ lệ chính xác cao 100% khi không có lỗi nào.

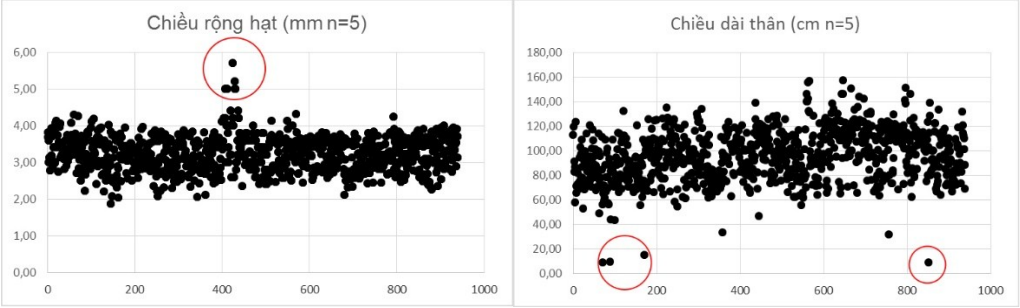
#### 3.1.2 Nhận diện, xử lý phần tử ngoại lai (outliers) và giảm thiểu nhiễu (noise data)

- Các phần tử ngoại lai (Outliers) có ảnh hưởng lớn đến độ chính xác của các mô hình dự đoán. Phát hiện và xử lý các điểm ngoại lai là một bước quan trọng trong quá trình chuẩn bị dữ liệu cho mô hình dự đoán. Những phần tử ngoại lai (đối tượng) này không tuân theo đặc tính/ hành vi chung của tập dữ liệu (đối tượng). Các giá trị tương tự nhau sẽ được hiển thị theo một cụm, các giá trị nằm ngoài, bất thường chính là các phần từ ngoại lai (outiers) gây ra dữ liệu nhiễu (noisy data). Các phần tử ngoại lai này thường

xuất hiện trong các chỉ tiêu đánh giá định lượng như Chiều dài hạt, Chiều rộng hạt, Chiều cao cây, Số dảnh, Thời gian sinh trưởng…

- Quá trình phân tích dữ liệu mô tả đánh giá bằng phương pháp giảm thiểu nhiễu phân cụm (clustering) chúng tôi đã tìm ra được 2 chỉ tiêu có dữ liệu mà trong đó xuất hiện một số phần tử ngoại lai đó là chỉ tiêu *Chiều rộng hạt =0,5 cm - 0,57 cm* và chỉ tiêu *Độ dài thân < 30 cm*.

- 6 mẫu nguồn gen (GBVN017399 (Aroo ba trăng) GBVN017382 (Đha nang), GBVN017386 (Aroo đếêp Đha nang), GBVN017403 (Aroo đếêp prong), GBVN017404 (Aroo đếêp Arút), GBVN017405 (Aroo đếêp Adíp) ) có dữ liệu *Chiều rộng hạt >=0,5 cm* được đối chứng lại với seed file nguồn gen đã được mô tả đánh giá lại ; 02 mẫu nguồn gen có dữ liệu *Độ dài thân < 30 cm* đó là: GBVN017283 (Khẩu già zui); TEMP019134 (Tài lồ) được yêu cầu mô tả, đánh giá lại.




**Hình 1**: Các phần tử ngoại lai trong chỉ tiêu mô tả **Chiều rộng hạt, Chiều dài thân**

### 3.2.3. Nhận diện, xử lý dữ liệu bị thiếu (missing data)

- Dữ liệu bị thiếu (missing data) là dữ liệu không sẵn có khi cần sử dụng xuất hiện do khách quan (không tồn tại lúc nhập liệu, sự cố) hoặc chủ quan (tác nhân con người). Chúng tôi đã tiến hành xác định dữ liệu bị thiếu (missing data) trên nhóm dữ liệu mô tả, đánh giá nguồn gen và xử lý bằng cách đưa về hằng số "null" cho các giá trị này

**Bảng 2**: Thống kế số lương các trường dữ liệu bị thiếu trong dữ liệu MTĐG

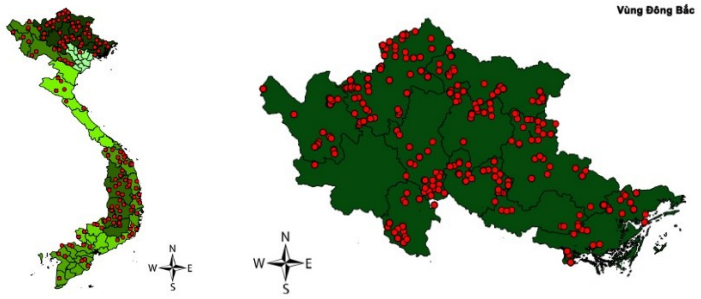
| ***Chỉ tiêu*** | **Màu phiến lá** | **Màu thìa lìa** | **Dạng thìa lìa** | **Màu cổ lá** | **Màu tai lá** | **Số dảnh hữu hiệu** | **Màu nhị cái** | **Màu ống rạ** | **Dạng bông** |
|---|---|---|---|---|---|---|---|---|---|
| ***Giá trị thiếu*** | 21 | 18 | 18 | 18 | 18 | 27 | 22 | 18 | 21 |
| ***Chỉ tiêu*** | **Độ thoát cổ bông** | **Trục bông** | **Râu** | **Màu mỏ hạt** | **Màu vỏ trấu** | **Độ phủ lông vỏ trấu** | **Màu mày hạt** | **Màu hạt gạo** | |
| ***Giá trị thiếu*** | 19 | 20 | 5 | 5 | 4 | 4 | 5 | 13 | |

- Việc xác định được số lượng các dữ liệu bị thiếu (missing data) giúp cho chúng tôi liệt kê danh sách những nguồn gen bị khuyết dữ liệu và có kế hoạch hoàn thiện dữ liệu trong các đợt nhân giống, mô tả đánh giá nguồn gen tiếp theo.

### 3.2 Ứng dụng Hệ thống thông tin địa lý (Geographic Information System- GIS) để phân tích, thống kê dữ liệu không gian (dữ liệu bản đồ) của các mẫu nguồn gen

- GIS từ lâu đã là công cụ hỗ trợ đắc lực để phân tích, hiển thị các thông tin liên quan tới vị trí địa lý của các đối tượng. Đối với dữ liệu của bảo tồn tài nguyên thực vật nông nghiệp, nếu chúng ta có một cơ sở dữ liệu nền tốt về vị trí địa lý, dữ liệu khí hậu, dữ liệu thổ nhưỡng… thì GIS sẽ giúp ích rất nhiều trong công tác mô phỏng, dự đoán.

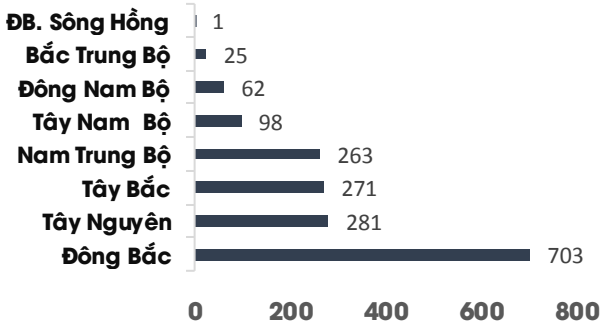
- Bước đầu ứng dụng phần mềm QGIS: dữ liệu thống kê đã cho thấy được sự phân bố đa dạng của 1.704 nguồn gen Lúa địa phương trải dài từ Bắc – Nam, một số vùng có hệ số đa dạng cao dựa trên số lượng nguồn gen như (Tây Bắc, Tây Nguyên, Đông Bắc), vùng có hệ số đa dạng thấp (ĐB Sông Hồng, Bắc Trung Bộ). Dữ liệu dạng bản đồ cũng cho cái nhìn khái quát về kết quả thu thập nguồn gen Lúa trong Dự án phát triển ngân hàng gen cây trồng quốc gia cũng như giúp lập kế hoạch trong các chương trình thu thập nguồn gen tại các vùng chưa được thu thập.




**Hình 2:** Thống kê phân bố nguồn gen Lúa thu thập sử dụng công cụ GIS

### 3.3 Ứng dụng phần mềm xử lý số liệu để phân tích, thống kê dữ liệu, Lai lịch, dữ liệu Mô tả đánh giá nguồn gen.

- Sử dụng các phần mềm xử lý thống kê dữ liệu Excel, SPSS chúng tôi đã phân tích thống kê dữ liệu Lai lịch của 1.704 nguồn gen và dữ liệu Mô tả, đánh giá của 940 nguồn gen theo nhiều hướng. Các kết quả của công tác phân tích, thống kê được trình bày theo các Hình, Bảng dưới đây:




**Hình 3**: Thống kê số lượng nguồn gen Lúa thu thập theo vùng sinh thái

**Bảng 3:** Thống kê số lượng Tỉnh/Huyện/Xã đã thu thập nguồn gen Lúa

| | Đông Bắc | Tây Nguyên | Tây Bắc | Nam Trung Bộ | Tây Nam Bộ | Đông Nam Bộ | Bắc Trung Bộ | ĐB. Sông Hồng | Tổng |
|---|---|---|---|---|---|---|---|---|---|
| **Tỉnh** | 11 | 5 | 4 | 5 | 7 | 3 | 2 | 1 | ***38*** |
| **Huyện** | 43 | 26 | 12 | 25 | 13 | 12 | 6 | 1 | ***133*** |
| **Xã** | 187 | 59 | 51 | 49 | 34 | 18 | 11 | 1 | ***410*** |

**Bảng 4:** Thống kê số lượng nguồn gen thu thập theo Dân tộc

| STT | Nhóm dân tộc | Số lượng nguồn gen | Tỷ lệ% | STT | Nhóm dân tộc | Số lượng nguồn gen | Tỷ lệ % |
|---|---|---|---|---|---|---|---|
| **1** | Kinh | 239 | 14.04 | **17** | Ê Đê | 13 | 0.76 |
| **2** | Dao | 222 | 13.04 | **18** | Kháng | 13 | 0.76 |
| **3** | H'Mông | 197 | 11.57 | **19** | Khơ Mú | 10 | 0.59 |
| **4** | Tày | 194 | 11.40 | **20** | Lô Lô | 7 | 0.41 |
| **5** | Nùng | 118 | 6.93 | **21** | Phù Lá | 7 | 0.41 |
| **6** | Xơ Đăng | 102 | 5.99 | **22** | Giáy | 6 | 0.35 |
| **7** | Thái | 89 | 5.23 | **23** | Cơ Ho | 5 | 0.29 |
| **8** | Cơ tu | 80 | 4.70 | **24** | La Chí | 5 | 0.29 |
| **9** | Gia rai | 73 | 4.29 | **25** | Hà Nhì | 5 | 0.29 |
| **10** | Ba Na | 67 | 3.94 | **26** | Lào | 4 | 0.24 |
| **11** | M'nông | 59 | 3.47 | **27** | Pà Thẻn | 4 | 0.24 |
| **12** | Mường | 51 | 3.00 | **28** | Hoa | 4 | 0.24 |
| **13** | Khmer | 43 | 2.53 | **29** | Co | 2 | 0.12 |
| **14** | Sán chay | 34 | 2.00 | **30** | Raglai | 2 | 0.12 |
| **15** | Giẻ Triêng | 28 | 1.65 | **31** | Chăm | 1 | 0.06 |
| **16** | Mạ | 17 | 1.00 | **32** | Sán Dìu | 1 | 0.06 |

**Bảng 5**: Thống kê các chỉ tiêu mô tả đánh giá ban đầu nguồn gen Lúa

| **Màu lá** | **Số lượng** | **Tỷ lệ %** | **Màu gốc bẹ lá** | **Số lượng** | **Tỷ lệ %** |
|---|---|---|---|---|---|
| 1– Xanh nhạt | 265 | 28,84 | 1– Xanh | 860 | 93,68 |
| 2– Xanh | 496 | 53,97 | 2– Có sọc tím | 35 | 3,81 |
| 3– Xanh đậm | 135 | 14,69 | 3– Tím nhạt | 19 | 2,07 |
| 4– Tím ở đỉnh | 0 | 0 | 4– Tím | 4 | 0,44 |
| 5– Tím ở mép lá | 14 | 1,52 | ***Tổng*** | ***918*** | |
| 6– Có đốm | 5 | 0,54 | | | |
| 7– Tím | 4 | 0,44 | | | |
| ***Tổng*** | ***919*** | | | | |
| **Màu cổ lá** | **Số lượng** | **Tỷ lệ %** | **Màu tai lá** | **Số lượng** | **Tỷ lệ %** |
| 1– Xanh nhạt | 716 | 77,66 | 1– Xanh nhạt | 841 | 91,21 |
| 2– Xanh | 158 | 17,14 | 2– Tím | 81 | 8,79 |
| 3– Tím | 48 | 5,21 | ***Tổng*** | ***922*** | |
| ***Tổng*** | ***922*** | | | | |
| **Màu nhụy** | **Số lượng** | **Tỷ lệ %** | **Màu sắc ống rạ** | **Số lượng** | **Tỷ lệ %** |

| | | | | | |
|---|---|---|---|---|---|
| 1– Trắng | 742 | 80,83 | 1– Xanh | 382 | 41,43 |
| 2– Xanh nhạt | 0 | 0 | 2– Vàng nhạt | 489 | 53,04 |
| 3– Vàng | 47 | 5,12 | 3– Sọc tím | 39 | 4,23 |
| 4– Tím nhạt | 76 | 8,28 | 4– Tím | 12 | 1,30 |
| 5– Tím | 53 | 5,77 | ***Tổng*** | ***922*** | |
| ***Tổng*** | ***918*** | | | | |
| **Dạng bông** | **Số lượng** | **Tỷ lệ %** | **Phân nhánh thứ cấp trên bông** | **Số lượng** | **Tỷ lệ %** |
| 1– Chụm | 99 | 10,77 | 1– Không | 7 | 0,76 |
| 5– Trung gian | 528 | 57,45 | 2– Nhẹ | 842 | 91,52 |
| 9– Mở | 292 | 31,77 | 3– Nặng | 67 | 7,28 |
| | | | 4– Đẻ cụm | 4 | 0,43 |
| | | | ***Tổng*** | ***920*** | |
| **Độ thoát cổ bông** | **Số lượng** | **Tỷ lệ %** | **Trục bông** | **Số lượng** | **Tỷ lệ %** |
| 1– Thoát hoàn | 698 | 75,79 | 1– Thẳng đứng | 7 | 0,76 |
| 3– Thoát trung bình | 162 | 17,59 | 2– Uốn xuống | 913 | 99,24 |
| 5– Vừa đúng cổ bông | 58 | 6,30 | Tổng | 920 | |
| 7– Thoát một | 2 | 0,22 | | | |
| 9– Không thoát được | 1 | 0,11 | | | |
| ***Tổng*** | ***921*** | | | | |
| **Râu** | **Số lượng** | **Tỷ lệ %** | **Màu râu** | **Số lượng** | **Tỷ lệ %** |
| 1– Không râu | 720 | 77,01 | 1– Vàng rơm | 42 | 19,09 |
| 3– Râu ngắn từng phần | 160 | 17,11 | 2– Vàng | 46 | 20,91 |
| 5– Râu ngắn toàn phần | 9 | 0,96 | 3– Nâu | 28 | 12,73 |
| 7– Râu dài từng | 34 | 3,64 | 4– Đỏ | 40 | 18,18 |
| 9– Râu dài toàn phần | 12 | 1,28 | 5– Tím | 48 | 21,82 |
| ***Tổng*** | ***935*** | | 6– Đen | 16 | 7,27 |
| | | | ***Tổng*** | ***220*** | |
| **Màu mỏ hạt** | **Số lượng** | **Tỷ lệ %** | **Màu vỏ trấu** | **Số lượng** | **Tỷ lệ %** |
| 1– Trắng | 20 | 2,14 | 1– Vàng rơm | 291 | 31,09 |
| 2– Vàng rơm | 375 | 40,11 | 2– Vàng hoặc khía vàng | 231 | 24,68 |
| 3– Nâu | 271 | 28,98 | 3– Đốm | 61 | 6,52 |
| 4– Đỏ | 26 | 2,78 | 4– Khía nâu | 197 | 21,05 |
| 5– Đỉnh đỏ | 4 | 0,43 | 5– Nâu | 33 | 3,53 |
| 6– Tím | 205 | 21,93 | 6– Hơi đỏ đến tím nhạt | 8 | 0,85 |
| 7– Đỉnh tím | 34 | 3,64 | 7– Đốm tím | 40 | 4,27 |
| ***Tổng*** | ***935*** | | 8– Khía tím | 57 | 6,09 |
| | | | 9– Tím | 14 | 1,50 |
| | | | 10– Đen | 4 | 0,43 |
| | | | ***Tổng*** | ***936*** | |
| **Màu mày hạt** | **Số lượng** | **Tỷ lệ %** | **Màu vỏ cám** | **Số lượng** | **Tỷ lệ %** |

| | | | | | |
|---|---|---|---|---|---|
| 1– Vàng rơm | 563 | 60,21 | 1– Trắng | 761 | 82,09 |
| 2– Vàng | 132 | 14,12 | 2– Nâu nhạt | 9 | 0,97 |
| 3– Đỏ | 111 | 11,87 | 3– Ánh nâu | 21 | 2,27 |
| 4– Tím | 129 | 13,80 | 4– Nâu | 10 | 1,08 |
| ***Tổng*** | ***935*** | | 5– Đỏ | 64 | 6,90 |
| | | | 6– Tím một phần | 16 | 1,73 |
| | | | 7– Tím | 46 | 4,96 |
| | | | ***Tổng*** | ***927*** | |

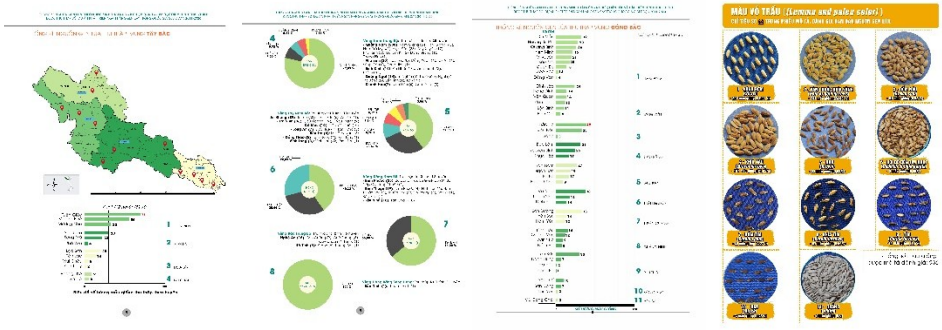
**Bảng 6:** Thống kê dữ liệu tính toán một số chỉ tiêu định lượng của Lúa

| | **D. thìa lìa** ***(mm, n=5)*** | **Số dảnh** | **TL.1000 hạt** ***(gr, n=3)*** | **Dài hạt** ***(mm, n=5)*** | **Rộng hạt** ***(mm, n=5)*** | **TG sinh trưởng** ***(ngày)*** |
|---|---|---|---|---|---|---|
| **Max** | 42.40 | 29.00 | 57.10 | 12.00 | 4.54 | 162.00 |
| **Min** | 1.00 | 1.30 | 10.00 | 6.04 | 1.88 | 95.00 |
| **Trung bình** | 16.66 | 6.66 | 28.62 | 8.76 | 3.24 | 127.18 |

### 3.4 Ứng dụng Infographics (Information graphic) để đồ họa trực quan thông tin, dữ liệu nguồn gen phục vụ công tác in ấn, xuất bản ấn phẩm thúc đẩy khai thác và sử dụng bền vững nguồn gen

- Infographic (thiết kế đồ họa thông tin) là kiểu thiết kế đồ họa chủ yếu dựa vào các hình ảnh trực quan để mô phỏng cho những dữ liệu thông tin, với thiết kế kiểu này người dùng dễ dàng thu thập dữ liệu một cách nhanh nhất nhờ các biểu tượng, các icon. Thời gian gần đây Infographic đã trở nên phổ biến cho những ý tưởng cho những thông tin phức tạp được trình bày trên nhiều bảng biểu, nhiều trang giấy.

- Thay vì tập hợp tin tức dạng text thì bây giờ chúng tôi sử dụng infographic để có thể thống kê 1 cách rõ ràng và chi tiết nhất, giúp cho các cán bộ nghiên cứu có thể hấp thụ và trao đổi nguồn thông tin dễ dàng hơn. Với những lợi ích của infographic, chúng tôi đã ứng dụng để xuất bản tài liệu ***"Thống kê nguồn gen Lúa theo vùng sinh thái nông nghiệp, dân tộc và đặc điểm hình thái chính được thu thập bởi Dự án phát triển ngân hàng gen cây trồng quốc gia giai đoạn 2011-2020".***



**Hình 4:** Infographic đồ họa trực quan dữ liệu thông tin được phân tích, thống kê

## IV. KẾT LUẬN VÀ ĐỀ NGHỊ

### 4.1 Kết luận

- Đã ứng dụng hiệu quả CNTT để làm sạch, xử lý phân tích, thống kê cơ sở dữ liệu lai lịch của 1.704 mẫu nguồn gen và mô tả đánh giá ban đầu của 940 mẫu nguồn gen lúa từ dự án;

- Ứng dụng thành công Hệ thống thông tin địa lý (GIS) để phân tích thống kê dữ liệu không gian các nguồn gen thu thập được từ dự án;

- Ứng dụng thành công Đồ họa trực quan hình ảnh (Infographic) để trình bày thông tin dữ liệu nguồn gen lúa từ dự án phục phục vụ in ấn, xuất bản. Đã xuất bản được 01 ấn phẩm thống kê nguồn gen Lúa phục vụ khai thác sử dụng nguồn gen;

- Dữ liệu được phân tích, thống kê theo nhiều hướng giúp ích cho các nhà nghiên cứu có cái nhìn đa chiều về công tác bảo tồn tài nguyên thực vật nông nghiệp.

### 4.2. Đề nghị

- Cần tiếp tục ứng dụng CNTT để phân tích, thống kê dữ liệu của các Loại cây khác, nhóm cây khác trong toàn hệ thống Bảo tồn nguồn gen thực vật nông nghiệp.

- Tiếp tục ứng dụng những khái niệm CNTT mới trong công tác tư liệu hóa thông tin nguồn gen.

## TÀI LIỆU THAM KHẢO


1. Hà Quang Thụy, Phan Xuân Hiếu, Đoàn Sơn, Nguyễn Trí Thành, Nguyễn Thu Trang, Nguyễn Cẩm Tú, **Giáo trình Khai phá dữ liệu Web**, NXB Giáo dục, 2009
2. TS. Nguyễn Minh Tuấn, Hà Trọng Quang, **Giáo trình Xử lý dữ liệu nghiên cứu với SPSS FOR WINDOW**, Trường ĐH Công nghiệp TP.HCM
3. https://ongxuanhong.wordpress.com/2016/01/31/lay-va-lam-sach-du-lieu-xu-ly-du-lieu-ngoai-lai-outliers/
4. Khai thác dữ liệu & ứng dung (Data Mining) (http://tailieu.vn/doc/data-mining-and-application-qui-trinh-chuan-bi-du-lieu-723931.html)
5. Tài liệu xử lý thống kê bằng Excel (http://tailieu.vn/doc/xu-ly-thong-ke-bang-excel-365594.html)
6. Tài liệu QGIS (http://www.qgistutorials.com/vi/)
7. Cơ sở dữ liệu 54 dân tộc Việt Nam (http://www.cema.gov.vn/gioi-thieu/cong-dong-54-dan-toc.htm
8. https://en.wikipedia.org/wiki/Infographic
9. Robert Nisbet, John Elder, Gary Miner**, Handbook of Statistical Analysis and Data Mining Applications**, Elsevier Inc, 2009.